

Diffusion Approximations for Queueing Systems with Customer Abandonment

Junfei Huang^{*}, Hanqin Zhang^{*}, and Jiheng Zhang[†]

^{*}National University of Singapore

[†]The Hong Kong University of Science and Technology

August 29, 2012

Abstract

This paper studies the diffusion approximations for queueing systems with customer abandonment. As queueing systems arise from various applications and the abandonment phenomenon has been recognized to be significant, we develop the diffusion analysis for such systems within a general framework for modeling the abandonment. In light of the recent work by Dai and He (2010), our first step is to show that, under an appropriate condition, the diffusion-scaled cumulative number of customers who have abandoned is asymptotically close to a function of the diffusion-scaled queue length process. Applying this asymptotic relationship, the diffusion approximations for such systems are established in a wide range of heavy traffic regimes, including the traditional heavy traffic regime, the non-degenerate slowdown regime and the Halfin-Whitt regime. The key step is to develop mappings with nice properties that can characterize the system dynamics by connecting the primitive processes such as the arrival, service and abandonment processes to the processes underlying the systems such as the queue length or the total head count.

1 Introduction

Customer abandonment has been extensively studied in various queueing models, motivated by their applications in different service systems. Many studies focus on diffusion analysis of the queueing processes in heavy traffic regimes where the arrival rate and service capacity increase to infinity while staying roughly balanced. The asymptotic studies in such regimes indeed make sense since the utilization is close to one in many real service systems which are often quite large. This paper focuses on providing a unified approach to the diffusion analysis for the different scalings of the patience time distribution, and for the various heavy traffic regimes. The unified approach applies not only to several classical models, which have been studied before, but also to a few cases which have not yet been explored in the literature. The models considered in this paper can be generally described as queueing systems each with a single customer class and a single server pool, denoted by $G/GI/N_n + GI$. Here, N_n is the number of homogeneous servers in the server pool. We will explain in the following section how to scale N_n in the general framework of heavy traffic regimes.

1.1 On heavy traffic regimes

The queueing systems with complex dynamics are usually studied in a heavy traffic regime. Loosely speaking, a heavy traffic regime is designed such that a significant number of events (customer arrivals, abandonments and service completions) can happen in a reasonable amount

of time. This allows the functional law of large numbers or functional central limit theorem (FCLT) to be used to approximate the stochastic processes underlying the queueing system. This kind of probabilistic characterizations can help elucidate how the systems work, and possibly provide insights for optimal design and control of such systems.

The general setting for a heavy traffic regime is to consider a sequence of systems indexed by n . Denote by λ^n the arrival rate to the n th system, and assume that

$$\frac{\lambda^n}{n} \rightarrow \mu, \quad \text{as } n \rightarrow \infty,$$

for some constant $\mu > 0$. So the customer arrival rate increases linearly to infinity in the heavy traffic regime. To balance the increasing demand, the service capacity must also increase at the same rate. Assume that the number of servers $N_n = n^\alpha$ for some parameter $\alpha \in [0, 1]$, and each server can process customers at rate $\mu^n = n^{1-\alpha}\mu$. In this way, the total service rate amounts to $n\mu$ and just balances out the arrivals. On a more detailed level, assume that for some $\beta \in \mathbb{R}$,

$$\sqrt{n}\left(\frac{\lambda^n}{n\mu} - 1\right) \rightarrow \beta, \quad \text{as } n \rightarrow \infty.$$

This roughly means that there is an imbalance of order \sqrt{n} between the service capacity and the arrival rate when the magnitude of the arrival rate is of order n . We are mainly interested in how the stochastic process $X^n(t)$, which counts the number of customers in the system at time t , oscillates around N_n .

The above heavy traffic framework with parameter α was first set out by Atar (2012), in an effort to bridge the so-called conventional heavy traffic regime and the many-server or the Halfin-Whitt heavy traffic regime. The conventional heavy traffic regime is mainly for single server queues, i.e., $N_n = 1$ or equivalently $\alpha = 0$. This regime dates back to Kingman (1961, 1962), with the FCLT point of view being developed by Iglehart and Whitt (1970a,b). For surveys see Glynn (1990) and Whitt (2002). In this regime, the diffusion limit can normally be characterized using a certain function of Brownian motions. The methodology and results can be extended to a general network setting, but with quite substantial efforts. See, for example Reiman (1984) and Williams (1998), where high-dimensional Brownian motions are involved.

Considering the other extreme by letting $\alpha = 1$ yields the Halfin-Whitt regime in which only the number of servers increases with n while the service rate of each server is kept fixed. The regime was initiated in the seminal work of Halfin and Whitt (1981), where service times were assumed to follow the exponential distribution. Since then there has been much effort expended to solve the problem by relaxing the assumption of exponentially distributed service times. Puhalskii and Reiman (2000) extended the diffusion analysis to allow the service time to follow phase-type distributions, and Whitt (2005) studied the same model with a finite buffer by assuming the service time follows a special class of service time distribution, denoted by H_2^* , which is a mixture of an exponential distribution and a unit point mass at 0. Both of these studies were conducted within a Markovian framework by taking advantage of the memoryless property of exponential random variables. A breakthrough was made by Reed (2009) and Puhalskii and Reed (2010), who proposed an innovative representation of the system dynamics and a regulator mapping to establish the diffusion limit in the Halfin-Whitt regime. The main idea is to connect the infinite server queue to the many-server queue using a regulator mapping. In a subsequent work, Reed (2007) described a separate approach which views the many-server queue from the idle time perspective and leads to an equivalent diffusion limit. Recently, Kaspi and Ramanan (2011b) characterized the diffusion limit of the many-server queue using a stochastic partial differential equation. The approach is based on the use of a measure-valued process to keep track of the residual service times of the customers in the system at any time. The framework of measure-valued processes was first developed in an earlier work of Kaspi and Ramanan (2011a), which analyzed the fluid limit of the system. Jelenković et al.

(2004) demonstrated the convergence of the steady state distribution of the model with deterministic service times. Their analysis takes advantage of the fact that there is no “overtaking” (meaning customers depart in the same order as they arrive) in many-server queues when the service time is deterministic. Gamarnik and Momčilović (2008) analyzed the stationary value of the queue length process assuming a lattice-valued service time distribution.

For both the conventional and Halfin-Whitt heavy traffic regimes, the arrival and total service rates are of order $O(n)$. Existing research shows that the queue length scales as $O(n^{1/2})$, thus the expected delay scales as $O(n^{-1/2})$. The difference between the two regimes lies in how each of them achieves the total service rate of order $O(n)$. In the conventional one the number of servers is fixed at one, while the service rate of the single server scales as $O(n)$, resulting in the service time scaling as $O(n^{-1})$. On the contrary, in the Halfin-Whitt regime the number of servers scales as $O(n)$ while the service rate of an individual server is fixed at order $O(1)$, i.e., the service time scales as $O(1)$. The different scalings originate from different applications. The conventional regime often applies to manufacturing or data package transmission over the Internet where the service rates of machines can be quite fast. The Halfin-Whitt regime is typically applied to service systems where humans are involved, because it is easier to increase the number of servers to achieve the total service rate of order $O(n)$ than to make each individual server work faster.

As in the work of Atar (2012), the *slowdown* is defined as the ratio between the sojourn time and the service time experienced by a typical customer. The above discussion shows that the slowdown degenerates in heavy traffic to ∞ and 1 in the conventional and Halfin-Whitt regimes, respectively. In fact, as pointed out by Atar (2012), the degeneration occurs for any $\alpha \in [0, 1]$ except $\alpha = 1/2$. The heavy traffic regime with parameter $\alpha = 1/2$ is therefore referred to as the *non-degenerate slowdown* (NDS) regime. The same regime, where delay and service times remain comparable, was also considered by Mandelbaum (2003), Whitt (2003) and Gurvich (2004). Readers are referred to Atar (2012) for very interesting examples where this regime becomes useful.

1.2 On customer abandonment

On top of the basic model we can add abandonment to allow customers to leave the system during waiting. This is motivated by the fact that in applications outstanding orders may be canceled in manufacture industries, and data packets have a high probability to be dropped if the waiting time is too long in a transmission channel, and customers can easily hang up at a call center after waiting for a while. Abandonment is modeled by assuming each customer has a patience time, which is a random variable. A customer abandons the system once his waiting time exceeds his patience time. The study of customer abandonment dates back to Palm (1937), who identified the impatient behavior of telephone switchboard customers.

In the current literature, there are two main streams of studies on abandonment, distinguished by different scalings of the patience time. The first one keeps the patience time distribution fixed in a heavy traffic regime. This stream can be further classified depending on the heavy traffic regime assumed. In the conventional heavy traffic regime, Ward and Glynn (2003) identified the diffusion limit as a reflected Ornstein-Uhlenbeck process for the $M/M/1 + M$ model. Later, Ward and Glynn (2005) extended the result to the general $G/GI/1 + GI$ model. In the Halfin-Whitt regime, Garnett et al. (2002) obtained the diffusion limit for the $M/M/n + M$ model. Recently, Dai et al. (2010) extended the diffusion analysis to a more general $G/PH/n + GI$ model by applying a general continuous map to both the fluid and diffusion-scaled processes and the random-time-change theorem. At the same time, Mandelbaum and Momčilović (2012) derived diffusion approximations by connecting the $G/GI/n + GI$ queue to the $G/GI/n$ queue, building on the work of Reed (2009), and identifying a relation between the queue length and

the offered waiting time processes. In the NDS regime, Atar (2012) established the diffusion approximation for the model with Poisson arrivals, and exponential service and patience times. Results in all of the above studies share the common feature that only the density of the patience time distribution at the origin plays a role in the diffusion limit. Intuitively, the reason is that the expected delay being of order $O(n^{-1/2})$ causes the abandonment behavior to be dominated by the shape of the patience time distribution between 0 and $O(n^{-1/2})$. Thus, taking the limit $n \rightarrow \infty$ loses all the information about the patience time distribution except for the density at 0.

In the other stream of the literature, the patience time distribution is instead scaled by using the hazard rate. For the n th system, the hazard rate function is defined by

$$h^n(x) = h(\sqrt{n}z),$$

as introduced by Reed and Ward (2008). The motivation behind this scaling is to preserve more information about the patience time distribution rather than just the density at a single point. For a more general scaling, see Lee and Weerasinghe (2011). As was pointed out by Zeltyn and Mandelbaum (2005) based on a statistical study of call center data, the estimate for the hazard rate function at a single point turns out to be unstable. Reed and Ward (2008) obtained the diffusion approximations for both the offered waiting time process and the queue length process for the $G/GI/1 + GI$ model in the conventional heavy traffic regime. Their approach was to use a non-linear generalized regulator mapping to establish the weak convergence results. Recently, Reed and Tezcan (2011) applied the same hazard rate scaling to study the diffusion limit for the queue length process of the $G/M/n + GI$ model by proving the asymptotic equivalence between the queue length process and the virtual waiting time process. In both studies the stationary distribution of the limiting processes was analyzed to obtain approximations for various performance measures. Extensive numerical experiments by Reed and Tezcan (2011) showed that the approximations involving the entire hazard rate function generally outperform those that only rely on the density at the origin. In particular, when the distribution density of the patience time changes rapidly near the origin, using only the density at zero to estimate the drift term of the diffusion limit yields a sizable error. See Section 4 of Reed and Tezcan (2011) for a detailed explanation.

Table 1.1 summarizes the existing studies on the diffusion analysis of queueing systems at the process level by classifying them into three heavy traffic regimes and two scalings of the patience time distribution. In addition to the research focusing on process level analysis,

	No Scaling	Hazard Rate Scaling
Conventional	Ward and Glynn (2003) $M/M/1 + M$ Ward and Glynn (2005) $G/GI/1 + GI$ Lee and Weerasinghe (2011) $G/GI/1 + GI$	Reed and Ward (2008) $G/GI/1 + GI$ Lee and Weerasinghe (2011) $G/GI/1 + GI$
NDS	Atar (2012) $M/M/\sqrt{n} + M$	
Halfin-Whitt	Garnett et al. (2002) $M/M/n + M$ Dai et al. (2010) $G/PH/n + GI$ Mandelbaum and Momčilović (2012) $G/GI/n + GI$	Reed and Tezcan (2011) $G/M/n + GI$

Table 1.1: Diffusion Approximations for Systems with Abandonment

there have been quite a few studies focusing on the steady state for such systems in heavy traffic. Most of them considered the case where the patience time distribution is not scaled. For example, Zeltyn and Mandelbaum (2005) studied various performance measures of the $M/M/n + GI$ queue based on Baccelli and Hebuterne (1981), which analyzed the $M/GI/1 + GI$

queue. Boxma and de Waal (1994) developed several approximations to the probability of abandonment for the $M/GI/n + GI$ queue and performed some simulation tests. A more detailed survey is presented in Zeltyn and Mandelbaum (2005), which studied the steady state of the $M/M/n + G$ queue.

1.3 Our approach

In this paper, we aim to provide a unified approach to studying customer abandonment for different scalings of the patience time distribution in the various heavy traffic regimes parameterized by α . We first provide a general scaling of the patience time distribution $F^n(\cdot)$ for the n th system in a heavy traffic regime by assuming

$$\sqrt{n}F^n\left(\frac{x}{\sqrt{n}}\right) \rightarrow f(x), \quad \text{as } n \rightarrow \infty. \quad (1.1)$$

This asymptotic framework can unify the cases of both no scaling and hazard rate scaling. If there is no scaling, i.e., $F^n(x) = F(x)$ with $F(0) = 0$ and $F'(0) = \alpha$, then $f(x) = \alpha x$. With hazard rate scaling, the distribution function $F^n(x) = 1 - \exp(-\int_0^x h(\sqrt{n}t)dt)$ for some hazard rate function $h(\cdot)$. In this case, $f(x) = \int_0^x h(t)dt$. As we can see from the existing studies, the function $f(x)$ given by (1.1) can characterize the customer abandonment in the diffusion limit for both cases. The intuition of (1.1) is based on the observation that customers' waiting times are of order $O(n^{-1/2})$ in the class of heavy traffic regimes we are focusing on. Thus no matter how the patience time distribution is scaled, what is required is a meaningful limit for the behavior of the patience time distribution $F^n(\cdot)$ on the interval from 0 to $O(n^{-1/2})$. That is why we scale the parameter inside of $F^n(\cdot)$ by \sqrt{n} . What the function $f(x)/x$ captures is the asymptotic behavior of

$$\frac{F^n(x/\sqrt{n}) - F^n(0)}{x/\sqrt{n}},$$

which turns out to be the density at the origin if there is no scaling, and a function involving the hazard rate if we use hazard rate scaling. In a more general sense, other scaling methods may be devised to emphasize different aspects of the patience time distribution depending on the application. It is worth pointing out that the patience time distribution is not allowed to have an atom at the origin, as has been required by the existing studies, meaning that all customers have some degree of patience. In the case where there is a non-trivial portion of the customers who would leave immediately upon arrival if there is a queue, the study will require some other techniques which are beyond the scope of this study.

Based on the general framework (1.1) for the scaling of the patience time distribution, we first identify an asymptotic sample-path relationship between the customer abandonment process and the queue length process in Theorem 2.1 in the next section. When (1.1) is specialized to the case without scaling, our result reduces to that of Dai and He (2010). The relationship is established by connecting the abandonment process to the offered waiting time using the patience time distribution. The offered or the virtual waiting time can then be related to the queue length using a generalization of Little's law. The challenge of such study in the general scaling framework (1.1) is that the queue length processes are required to be tight, while only stochastic boundedness is needed for the case without scaling as in Dai and He (2010). Tightness, in particular the modulus of continuity (2.6), is usually difficult to verify, thus making Theorem 2.1 less useful in the diffusion analysis. To solve this issue, we develop a useful tool, Corollary 2.1, which establishes the tightness of the abandonment processes only based on the stochastic boundedness of the queue length processes. We later design "reflection" mappings which take the abandonment processes together with other processes (such as the arrival process), which can be shown to have good properties, as the input. With these nice properties in hand, the queue length (or head count) process, which is the output of those

mappings, are thus tight. This enables us to apply Theorem 2.1 in the diffusion analysis for the head count and queue length processes.

The service completion process is also an element of the input to the “reflection” mappings, so the next step is to characterize the service completion process using the information of the system size. When the service time is exponentially distributed, we have a unified approach for all the heavy traffic regimes parameterized by $\alpha \in [0, 1)$. When the service time distribution is general, we have to tailor the analysis to the specific regime where $\alpha = 1$, since the generality of the distribution not only requires different methods, but may also lead to different results in different regimes. So far, we have not been able to figure out the service completion process for a general service time distribution when $\alpha \in [0, 1)$. Thus, we leave it open to future research.

The above two steps enable us to transform the balance equation for the system into an equation which only involves the status of the system. To characterize the diffusion limit of the head count process, we develop a general approach to represent the head count process as the mapping of an input process involving the arrival, service completion and abandonment processes. Our analysis reveals that the heavy traffic regimes parameterized by $\alpha \in [0, 1)$ and $\alpha = 1$ are intrinsically different from the mapping point of view. So we divide the analysis into two cases. For the case with $\alpha \in [0, 1)$, a sequence of “reflection” mappings is developed in Lemma 4.1. The insight from these mappings shows that diffusion-scaled total head count processes cannot be negative in the limit, though for each fixed n th system it can be below 0 (meaning there are idling servers). Despite the limitation of analyzing the service completion process based on the exponential service time distribution, our “reflection” mapping approach still applies in general once the service completion process can be characterized for general service times. For the case with $\alpha = 1$, we are able to obtain the diffusion limit for the system when the service time is generally distributed. In this case, we also develop a regulator mapping (see Lemma 5.1), which is essentially an extension to the one in Section 4 of Reed (2007). However, it is different from that one and is new to the best of our knowledge. Summarizing the above, our approach to the diffusion approximation is depicted in Figure 1.1.

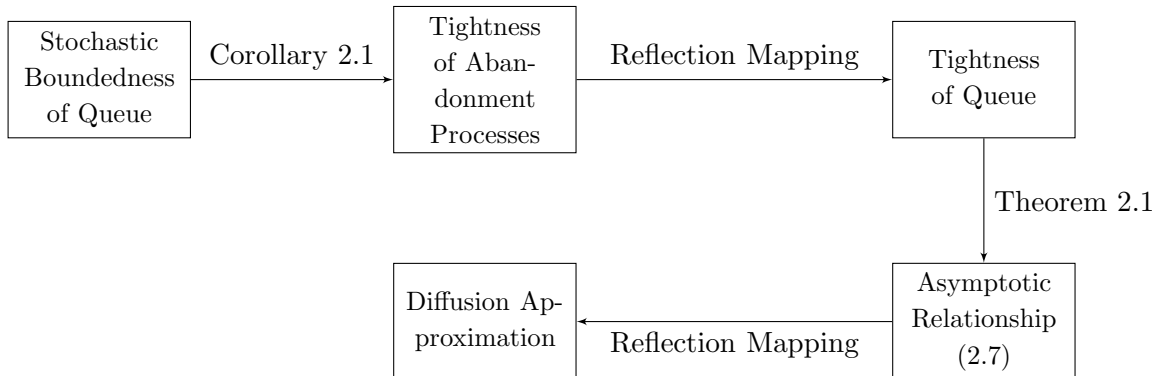


Figure 1.1: The approach to the diffusion approximation.

The rest of this paper is organized as follows. We propose a unified framework for studying the abandonment in Section 2. A three-step procedure (Propositions 2.1–2.3) is outlined in Section 2.1 and the complete proofs are given in Section 2.2. In Section 3, we present the results on diffusion approximations for many-server systems with abandonment where the number of servers $N_n = n^\alpha$ by making some regularity assumptions. The analysis is then divided into two cases. Section 4 analyzes the systems in the heavy traffic regimes where $\alpha \in [0, 1)$ and Section 5 analyzes them in the regime where $\alpha = 1$. The mapping for $\alpha = 1$ is studied in the Appendices.

To conclude this section, we introduce some notation and definitions which will be used throughout the paper. All random variables and processes are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ unless otherwise specified. We denote by \mathbb{Z}_+ , \mathbb{R} and \mathbb{R}_+ the sets of positive

integers, real numbers and nonnegative numbers, respectively. The space of RCLL (right continuous with left limits) functions on \mathbb{R}_+ taking values in \mathbb{R} is denoted by $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$, and the subspace of the continuous functions in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ is denoted by $\mathbf{C}(\mathbb{R}_+, \mathbb{R})$. For $g \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$, $g(t-)$ represents its left limit at $t > 0$, and the uniform norm of $g(\cdot)$ on the interval $[a, b]$ is defined by

$$\|g\|_{[a,b]} = \sup_{t \in [a,b]} |g(t)| \quad \text{with } \|g\|_{[0,b]} \text{ abbreviated to } \|g\|_b.$$

For a sequence of random elements $\{X^n, n \in \mathbb{Z}_+\}$ taking values in a metric space, we write $X^n \Rightarrow X$ to denote the convergence of X^n to X in distribution. The space $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ is assumed to be endowed with the Skorohod J_1 -topology (see Billingsley (1999)). For a probability distribution function $G(\cdot)$ defined on \mathbb{R}_+ , $G^c(\cdot) = 1 - G(\cdot)$. For $a \in \mathbb{R}$, $a^+ = \max\{a, 0\}$, $a^- = \max\{-a, 0\}$ and $\lfloor a \rfloor$ is the largest integer not greater than a . We use $\mathbf{1}_S$ to denote the indicator function of set $S \subset \Omega$.

2 The Framework for Customer Abandonment

Consider a sequence of single station queueing systems indexed by $n \in \mathbb{Z}_+$. Denote $Q^n(t)$ as the number of customers in the queue at time t , and $G^n(t)$ as the number of customers who have abandoned the queue by time t , in the n th system. Clearly, $G^n(0) = 0$ and $Q^n(0)$ is the number of customers who are waiting in queue at time zero. Define the diffusion-scaled processes $\tilde{Q}^n = \{\tilde{Q}^n(t) : t \geq 0\}$ and $\tilde{G}^n = \{\tilde{G}^n(t) : t \geq 0\}$ by

$$\tilde{Q}^n(t) = \frac{Q^n(t)}{\sqrt{n}}, \quad \tilde{G}^n(t) = \frac{G^n(t)}{\sqrt{n}}.$$

Our objective in this section is to prove an asymptotic relationship (Theorem 2.1) between \tilde{Q}^n and \tilde{G}^n under appropriate assumptions.

Let $E^n(t)$ denote the number of arrivals by time t in the n th system, and define the diffusion-scaled arrival process $\tilde{E}^n = \{\tilde{E}^n(t) : t \geq 0\}$ by

$$\tilde{E}^n(t) = \frac{E^n(t) - \lambda^n t}{\sqrt{n}}.$$

We assume that there is a sequence of constants $\{\lambda^n, n \in \mathbb{Z}_+\}$ with

$$\lim_{n \rightarrow \infty} \frac{\lambda^n}{n} = \mu > 0 \tag{2.1}$$

such that

$$\tilde{E}^n \Rightarrow \tilde{E} \quad \text{as } n \rightarrow \infty, \tag{2.2}$$

for some process $\tilde{E} = \{\tilde{E}(t) : t \geq 0\} \in \mathbf{C}(\mathbb{R}_+, \mathbb{R})$. We now introduce a sequence of i.i.d. random variables $\{\gamma_i^n, i \in \mathbb{Z}_+\}$, where γ_i^n is interpreted as the patience time of the i th arriving customer in the n th system. A customer waiting in the system will leave the system without receiving service once his patience time exhausts. The sequence $\{\gamma_i^n, i \in \mathbb{Z}_+\}$ is independent of the arrival process E^n for each n . In this paper, we assume for technical convenience that the patience times of the customers who are initially in the system are infinite, i.e., the initial customers in the queue are infinitely patient. We denote the patience time distribution as $F^n(\cdot)$ and assume that for $x \geq 0$,

$$\sqrt{n}F^n\left(\frac{x}{\sqrt{n}}\right) \rightarrow f(x), \quad \text{as } n \rightarrow \infty, \tag{2.3}$$

where $f(\cdot)$ is a nondecreasing and locally Lipschitz continuous function, i.e., for any $T \geq 0$, there is a constant Λ_T such that for all $x, y \in [0, T]$,

$$|f(x) - f(y)| \leq \Lambda_T |x - y|. \tag{2.4}$$

As pointed out in Section 1, there are two special cases corresponding to different approaches of scaling the patience time distribution:

- *No scaling.* Let $F^n(x) = F(x)$ where F is a probability distribution function with $F(0) = 0$ and $F'(0) = \alpha$. In this case, $f(x) = \alpha x$.
- *Hazard rate scaling.* Let $F^n(x) = 1 - \exp(-\int_0^x h(\sqrt{nt})dt)$ for some locally bounded hazard rate function $h(\cdot)$. In this case, $f(x) = \int_0^x h(t)dt$.

In order to obtain the asymptotic relationship (Theorem 2.1), the key assumption is that the sequence of diffusion-scaled queue length processes $\{\tilde{Q}^n, n \in \mathbb{Z}_+\}$ is C -tight. Namely the sequence is stochastically bounded, i.e., for each $T > 0$,

$$\lim_{\Gamma \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \tilde{Q}^n(t) > \Gamma \right\} = 0, \quad (2.5)$$

and the modulus of continuity is asymptotically small, i.e., for any $\varepsilon > 0$,

$$\lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{s, t \in [0, T], |s-t| < \delta} |\tilde{Q}^n(s) - \tilde{Q}^n(t)| > \varepsilon \right\} = 0. \quad (2.6)$$

Theorem 2.1. *If a sequence of $G/G/N_n + GI$ queues satisfies (2.1)–(2.6), then for each $T > 0$,*

$$\sup_{0 \leq t \leq T} \left| \tilde{G}^n(t) - \mu \int_0^t f\left(\frac{1}{\mu} \tilde{Q}^n(s)\right) ds \right| \Rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (2.7)$$

The statement of the result is the same as Theorem 2.1 of Dai and He (2010) who studied the same queueing process when patience time is not scaled. In other words, the function $f(x) = \alpha x$ where $\alpha = F'(0)$. However, due to the general framework (2.3) for treating patience time distribution, we need a stronger condition. The additional assumption we need is (2.6), which is mainly used to deal with the nonlinearity of the function f .

2.1 Outline of the Proof

The proof of Theorem 2.1 is based on three properties of such queueing systems, namely Propositions 2.1–2.3, which are of independent interest themselves. We will first state these properties and then apply them to prove Theorem 2.1. The proofs of the three propositions are postponed to Section 2.2.

Following Dai and He (2010), we introduce two notions. The first one is the *offered waiting time* ω_i^n , which denotes the time that the i th arriving customer in the n th system after time 0 has to wait before receiving service for each $i \geq 1$. When $Q^n(0) > 0$, we index the initial customer in the queue by $0, -1, \dots, -Q^n(0) + 1$, with customer $-Q^n(0) + 1$ being the first one in the queue. Each ω_i^n denotes the remaining waiting time of the i th customer for $i = -Q^n(0) + 1, \dots, 0$. The second notion is the *virtual waiting time* $\omega^n(t)$, which is the amount of time a hypothetical customer with infinite patience would have to wait before receiving service had he arrived at time t in the n th system. We introduce the diffusion-scaled virtual waiting time process $\tilde{\omega}^n = \{\tilde{\omega}^n(t) : t \geq 0\}$ as

$$\tilde{\omega}^n(t) = \sqrt{n} \omega^n(t).$$

The first property of interest is the stochastic boundedness of the virtual waiting time.

Proposition 2.1. *Under assumptions (2.1)–(2.5), the sequence of the scaled virtual waiting times $\{\tilde{\omega}^n, n \in \mathbb{Z}_+\}$ is also stochastically bounded.*

The second proposition reveals an asymptotic relationship between the abandonment process and the offered waiting time.

Proposition 2.2. *Under assumptions (2.1)–(2.5), for each $T > 0$,*

$$\sup_{0 \leq t \leq T} \left| \tilde{G}^n(t) - \frac{1}{\sqrt{n}} \sum_{j=1}^{E^n(t)} F^n(\omega_j^n) \right| \Rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

Note that neither of the above two propositions needs the modulus of continuity to vanish asymptotically as in (2.6). The next proposition establishes the relationship between the virtual waiting time and the queue length. For this one, condition (2.6) is required.

Proposition 2.3. *Under assumptions (2.1)–(2.6), for each $T > 0$,*

$$\sup_{0 \leq t \leq T} \left| \mu \tilde{\omega}^n(t) - \tilde{Q}^n(t) \right| \Rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

Remark 2.1. *By the above proposition and the triangle inequality, for any $s, t \in [0, T]$,*

$$\left| \tilde{\omega}^n(t) - \tilde{\omega}^n(s) \right| \leq 2 \sup_{0 \leq t \leq T} \left| \tilde{\omega}^n(t) - \frac{1}{\mu} \tilde{Q}^n(t) \right| + \frac{1}{\mu} \left| \tilde{Q}^n(t) - \tilde{Q}^n(s) \right|.$$

Thus, the C -tightness of $\{\tilde{Q}^n, n \in \mathbb{Z}_+\}$ implies that $\{\tilde{\omega}^n, n \in \mathbb{Z}_+\}$ is also C -tight.

Proof of Theorem 2.1. According to Proposition 2.2, it is enough to prove that as $n \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{E^n(t)} F^n(\omega_j^n) - \mu \int_0^t f\left(\frac{1}{\mu} \tilde{Q}^n(s)\right) ds \right| \Rightarrow 0. \quad (2.8)$$

After adding and subtracting a new term, we have

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{j=1}^{E^n(t)} F^n(\omega_j^n) - \mu \int_0^t f\left(\frac{1}{\mu} \tilde{Q}^n(s)\right) ds &= \frac{1}{\sqrt{n}} \sum_{j=1}^{E^n(t)} F^n(\omega_j^n) - \mu \int_0^t f(\tilde{\omega}^n(s)) ds \\ &\quad + \mu \int_0^t f(\tilde{\omega}^n(s)) ds - \mu \int_0^t f\left(\frac{1}{\mu} \tilde{Q}^n(s)\right) ds. \end{aligned}$$

Thus it is enough to prove that when $n \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \left| \int_0^t f(\tilde{\omega}^n(s)) ds - \int_0^t f\left(\frac{1}{\mu} \tilde{Q}^n(s)\right) ds \right| \Rightarrow 0, \quad (2.9)$$

$$\sup_{0 \leq t \leq T} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{E^n(t)} F^n(\omega_j^n) - \mu \int_0^t f(\tilde{\omega}^n(s)) ds \right| \Rightarrow 0. \quad (2.10)$$

We first prove (2.9). By assumption (2.5) and Proposition 2.1, for any $\varepsilon > 0$, there exists T_1 large enough such that for all large enough n ,

$$\mathbb{P}\left\{ \sup_{0 \leq t \leq T} \tilde{Q}^n(t) \geq T_1 \right\} + \mathbb{P}\left\{ \sup_{0 \leq t \leq T} \tilde{\omega}^n(t) \geq T_1 \right\} \leq \frac{\varepsilon}{2}. \quad (2.11)$$

For any $\delta > 0$, by the local Lipschitz continuity of $f(\cdot)$, we have

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{0 \leq t \leq T} \left| \int_0^t f(\tilde{\omega}^n(s)) ds - \int_0^t f\left(\frac{1}{\mu} \tilde{Q}^n(s)\right) ds \right| \geq \delta \right\} \\ &\leq \mathbb{P}\left\{ \sup_{0 \leq t \leq T} \tilde{Q}^n(t) \geq T_1 \right\} + \mathbb{P}\left\{ \sup_{0 \leq t \leq T} \tilde{\omega}^n(t) \geq T_1 \right\} + \mathbb{P}\left\{ \sup_{0 \leq t \leq T} \left| \tilde{\omega}^n(t) - \frac{1}{\mu} \tilde{Q}^n(t) \right| \geq \frac{\delta}{T\Lambda_{T_1}} \right\}. \end{aligned}$$

By Proposition 2.3, the third term on the right-hand side in the above can be less than $\varepsilon/2$ for all large enough n . Thus (2.9) is proved by (2.11).

Next, we prove (2.10). According to Lemma 3.2 of Dai and He (2010) and the monotonicity of the distribution function F^n ,

$$\int_0^t \sqrt{n} F^n\left(\frac{1}{\sqrt{n}} \tilde{\omega}^n(s-)\right) d\bar{E}^n(s) \leq \frac{1}{\sqrt{n}} \sum_{j=1}^{E^n(t)} F^n(\omega_j^n) \leq \int_0^t \sqrt{n} F^n\left(\frac{1}{\sqrt{n}} \tilde{\omega}^n(s)\right) d\bar{E}^n(s).$$

So it is sufficient to prove the following convergence as $n \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \left| \int_0^t \sqrt{n} F^n\left(\frac{1}{\sqrt{n}} \tilde{\omega}^n(s)\right) d\bar{E}^n(s) - \mu \int_0^t f(\tilde{\omega}^n(s)) ds \right| \Rightarrow 0, \quad (2.12a)$$

$$\sup_{0 \leq t \leq T} \left| \int_0^t \sqrt{n} F^n\left(\frac{1}{\sqrt{n}} \tilde{\omega}^n(s-)\right) d\bar{E}^n(s) - \mu \int_0^t f(\tilde{\omega}^n(s)) ds \right| \Rightarrow 0. \quad (2.12b)$$

We only prove (2.12a) since (2.12b) can be proved similarly. The idea is similar to the one proposed by Ward and Glynn (2005). Define the fluid-scaled arrival process $\bar{E}^n = \{\bar{E}^n(t) : t \geq 0\}$ by

$$\bar{E}^n(t) = \frac{E^n(t)}{n}.$$

Condition (2.2) implies that as $n \rightarrow \infty$,

$$\bar{E}^n \Rightarrow \bar{e} \quad \text{with} \quad \bar{e}(t) = \mu t. \quad (2.13)$$

By Remark 2.1, we have that $\{(\tilde{\omega}^n, \bar{E}^n), n \in \mathbb{Z}_+\}$ is C -tight. So for every convergent subsequence indexed by n_k ,

$$(\tilde{\omega}^{n_k}, \bar{E}^{n_k}) \Rightarrow (\tilde{\omega}, \bar{e}) \quad \text{as } n_k \rightarrow \infty,$$

for some process $\tilde{\omega} \in \mathbf{C}(\mathbb{R}_+, \mathbb{R})$. By the Skorohod representation theorem, there exists another probability space $(\check{\Omega}, \check{\mathcal{F}}, \check{\mathbb{P}})$, as well as a sequence of processes $(\check{\omega}^{n_k}, \check{E}^{n_k})$ and $(\check{\omega}, \bar{e})$ defined on it, such that

$$\begin{aligned} (\check{\omega}^{n_k}, \check{E}^{n_k}) &\stackrel{d}{=} (\tilde{\omega}^{n_k}, \bar{E}^{n_k}), \\ (\check{\omega}, \bar{e}) &\stackrel{d}{=} (\tilde{\omega}, \bar{e}), \end{aligned}$$

and with probability one, $\check{\omega}^{n_k}$ converges to $\check{\omega}$ in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ and \check{E}^{n_k} converges to \bar{e} in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$. By the continuous mapping theorem and (2.3), we also know that with probability one, both $f(\check{\omega}^{n_k})$ and $\sqrt{n_k} F^{n_k}(\frac{\check{\omega}^{n_k}}{\sqrt{n_k}})$ converge to $f(\check{\omega})$ as $n_k \rightarrow \infty$ in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$. By Lemma 8.3 of Dai and Dai (1999), we know that with probability one, as $n_k \rightarrow \infty$

$$\begin{aligned} \sup_{0 \leq t \leq T} \left| \int_0^t \sqrt{n_k} F^{n_k}\left(\frac{1}{\sqrt{n_k}} \check{\omega}^{n_k}(s)\right) d\check{E}^{n_k}(s) - \mu \int_0^t f(\check{\omega}(s)) ds \right| &\rightarrow 0, \\ \sup_{0 \leq t \leq T} \left| \int_0^t f(\check{\omega}^{n_k}(s)) ds - \int_0^t f(\check{\omega}(s)) ds \right| &\rightarrow 0. \end{aligned}$$

As a result, with probability one, we have as $n_k \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \left| \int_0^t \sqrt{n_k} F^{n_k}\left(\frac{1}{\sqrt{n_k}} \check{\omega}^{n_k}(s)\right) d\check{E}^{n_k}(s) - \mu \int_0^t f(\check{\omega}^{n_k}(s)) ds \right| \rightarrow 0. \quad (2.14)$$

Since $(\check{\omega}^{n_k}, \check{E}^{n_k}) \stackrel{d}{=} (\tilde{\omega}^{n_k}, \bar{E}^{n_k})$, we have

$$\begin{aligned} &\sqrt{n_k} \int_0^t F^{n_k}\left(\frac{1}{\sqrt{n_k}} \check{\omega}^{n_k}(s)\right) d\check{E}^{n_k}(s) - \mu \int_0^t f(\check{\omega}^{n_k}(s)) ds \\ &\stackrel{d}{=} \sqrt{n_k} \int_0^t F^{n_k}(\omega^{n_k}(s)) d\bar{E}^{n_k}(s) - \mu \int_0^t f(\tilde{\omega}^{n_k}(s)) ds. \end{aligned}$$

Hence (2.14) implies that as $k \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \left| \sqrt{n_k} \int_0^t F^{n_k}(\omega^{n_k}(s)) d\bar{E}^{n_k}(s) - \mu \int_0^t f(\tilde{\omega}^{n_k}(s)) ds \right| \Rightarrow 0.$$

Since the above convergence to zero holds for all convergent subsequences, (2.12a) is established. \square

2.2 Proofs of Propositions 2.1–2.3

In order to prove the propositions, we need the following lemma. For each $\delta > 0$, let

$$\tilde{L}_\delta^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \left(\mathbf{1}_{\{\gamma_i^n \leq \frac{\delta}{\sqrt{n}}\}} - F^n\left(\frac{\delta}{\sqrt{n}}\right) \right).$$

Lemma 2.1. *If (2.3) holds, then for any $\delta > 0$ and $T > 0$,*

$$\sup_{0 \leq t \leq T} |\tilde{L}_\delta^n(t)| \Rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

Proof. Denote $p_n = F^n\left(\frac{\delta}{\sqrt{n}}\right)$ and $X_{ni} = \mathbf{1}_{\{\gamma_i^n \leq \frac{\delta}{\sqrt{n}}\}} - p_n$. Then for any $\varepsilon > 0$,

$$\mathbb{P} \left\{ \sup_{0 \leq t \leq T} |\tilde{L}_\delta^n(t)| > 2\varepsilon \right\} = \mathbb{P} \left\{ \max_{1 \leq k \leq \lfloor nT \rfloor} |\tilde{L}_\delta^n\left(\frac{k}{n}\right)| > 2\varepsilon \right\}. \quad (2.15)$$

For $k = 0, \dots, \lfloor nT \rfloor - 1$, let

$$\tilde{L}_\delta^n(k, \lfloor nT \rfloor) = \frac{1}{\sqrt{n}} \sum_{i=k+1}^{\lfloor nT \rfloor} \left(\mathbf{1}_{\{\gamma_i^n \leq \frac{\delta}{\sqrt{n}}\}} - F^n\left(\frac{\delta}{\sqrt{n}}\right) \right).$$

Then

$$\begin{aligned} \mathbb{P} \left\{ |\tilde{L}_\delta^n(k, \lfloor nT \rfloor)| > \varepsilon \right\} &\leq \frac{1}{\varepsilon^2} \mathbb{E} \left(\tilde{L}_\delta^n(k, \lfloor nT \rfloor) \right)^2 \\ &= \frac{1}{\varepsilon^2 n} \sum_{i=k+1}^{\lfloor nT \rfloor} \mathbb{E} (X_{ni})^2 \\ &= \frac{1}{\varepsilon^2 n} \sum_{i=k+1}^{\lfloor nT \rfloor} \mathbb{E} \left(\mathbf{1}_{\{\gamma_i^n \leq \frac{\delta}{\sqrt{n}}\}} (1 - 2p_n) + p_n^2 \right) \\ &= \frac{1}{\varepsilon^2} \frac{\lfloor nT \rfloor - k}{n} p_n (1 - p_n). \end{aligned} \quad (2.16)$$

Hence,

$$\begin{aligned} \min_{1 \leq k \leq \lfloor nT \rfloor - 1} \mathbb{P} \left\{ |\tilde{L}_\delta^n(k, \lfloor nT \rfloor)| \leq \varepsilon \right\} &= 1 - \max_{1 \leq k \leq \lfloor nT \rfloor - 1} \mathbb{P} \left\{ |\tilde{L}_\delta^n(k, \lfloor nT \rfloor)| > \varepsilon \right\} \\ &\geq 1 - \frac{1}{\varepsilon^2} \frac{\lfloor nT \rfloor - 1}{n} p_n (1 - p_n). \end{aligned} \quad (2.17)$$

By the Ottaviani inequality (see Page 75 of Chow and Teicher (2003)) and (2.15), we have

$$\mathbb{P} \left\{ \sup_{0 \leq t \leq T} |\tilde{L}_\delta^n(t)| > 2\varepsilon \right\} \leq \frac{\mathbb{P} \left\{ |\tilde{L}_\delta^n(0, \lfloor nT \rfloor)| > \varepsilon \right\}}{\min_{1 \leq k \leq \lfloor nT \rfloor - 1} \mathbb{P} \left\{ |\tilde{L}_\delta^n(k, \lfloor nT \rfloor)| \leq \varepsilon \right\}}.$$

It directly follows from (2.16) with $k = 0$ and (2.17) that

$$\mathbb{P} \left\{ \sup_{0 \leq t \leq T} |\tilde{L}_\delta^n(t)| > 2\varepsilon \right\} \leq \left(\frac{1}{\varepsilon^2} \frac{\lfloor nT \rfloor}{n} p_n(1 - p_n) \right) / \left(1 - \frac{1}{\varepsilon^2} \frac{\lfloor nT \rfloor - 1}{n} p_n(1 - p_n) \right),$$

which converges to 0 as $n \rightarrow \infty$ according to assumption (2.3). \square

Proof of Proposition 2.1. The proof is similar to the proof of Proposition 4.4 in an earlier version of Dai and He (2010). First, by equation (33) on page 357 of Dai and He (2010), we have that for any $\Gamma \in (0, \tilde{\omega}^n(t))$,

$$E^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) - E^n(t) \leq Q^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) + \sum_{i=E^n(t)+1}^{E^n(t+\Gamma/\sqrt{n})} \mathbb{1}_{\{\gamma_i^n \leq \frac{\Gamma}{\sqrt{n}}\}}.$$

This implies that

$$\begin{aligned} \tilde{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) - \tilde{E}^n(t) + \frac{\lambda^n \Gamma}{n} &\leq \tilde{Q}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) + \tilde{L}_\Gamma^n \left(\bar{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) \right) - \tilde{L}_\Gamma^n (\bar{E}^n(t)) \\ &\quad + \sqrt{n} \cdot F^n \left(\frac{\Gamma}{\sqrt{n}} \right) \cdot \left(\bar{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) - \bar{E}^n(t) \right). \end{aligned}$$

Then

$$\begin{aligned} \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \tilde{\omega}^n(t) > \Gamma \right\} &\leq \mathbb{P} \left\{ \inf_{0 \leq t \leq T} \left[\tilde{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) - \tilde{E}^n(t) + \frac{\lambda^n \Gamma}{n} - \tilde{Q}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) \right. \right. \\ &\quad \left. \left. - \tilde{L}_\Gamma^n \left(\bar{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) \right) + \tilde{L}_\Gamma^n (\bar{E}^n(t)) \right. \right. \\ &\quad \left. \left. - \sqrt{n} \cdot F^n \left(\frac{\Gamma}{\sqrt{n}} \right) \cdot \left(\bar{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) - \bar{E}^n(t) \right) \right] \leq 0 \right\} \\ &\leq \mathbb{P} \left\{ \inf_{0 \leq t \leq T} \left[\tilde{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) - \tilde{E}^n(t) + \frac{\lambda^n \Gamma}{n} \right] \leq \frac{\mu \Gamma}{2} \right\} \\ &\quad + \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \tilde{Q}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) \geq \frac{\mu \Gamma}{12} \right\} \\ &\quad + \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \left| \tilde{L}_\Gamma^n \left(\bar{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) \right) - \tilde{L}_\Gamma^n (\bar{E}^n(t)) \right| \geq \frac{\mu \Gamma}{12} \right\} \\ &\quad + \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \sqrt{n} \cdot F^n \left(\frac{\Gamma}{\sqrt{n}} \right) \cdot \left(\bar{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) - \bar{E}^n(t) \right) \geq \frac{\mu \Gamma}{12} \right\}, \end{aligned}$$

where μ is given by (2.1). It follows from (2.1)–(2.2) that as $n \rightarrow \infty$,

$$\mathbb{P} \left\{ \inf_{0 \leq t \leq T} \left[\tilde{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) - \tilde{E}^n(t) + \frac{\lambda^n \Gamma}{n} \right] \leq \frac{\mu \Gamma}{2} \right\} \rightarrow 0, \quad (2.18)$$

Lemma 2.1 and (2.13) implies that as $n \rightarrow \infty$,

$$\mathbb{P} \left\{ \sup_{0 \leq t \leq T} \left| \tilde{L}_\Gamma^n \left(\bar{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) \right) - \tilde{L}_\Gamma^n (\bar{E}^n(t)) \right| \geq \frac{\mu \Gamma}{12} \right\} \rightarrow 0. \quad (2.19)$$

By (2.3) and (2.13), we have that as $n \rightarrow \infty$,

$$\mathbb{P} \left\{ \sup_{0 \leq t \leq T} \sqrt{n} \cdot F^n \left(\frac{\Gamma}{\sqrt{n}} \right) \cdot \left(\bar{E}^n \left(t + \frac{\Gamma}{\sqrt{n}} \right) - \bar{E}^n(t) \right) \geq \frac{\mu \Gamma}{12} \right\} \rightarrow 0. \quad (2.20)$$

Hence, the proposition follows from assumption (2.5) and (2.18)–(2.20). \square

Lemma 2.2. *Under assumptions (2.1)–(2.5),*

$$\lim_{\Gamma \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq i \leq E^n(T)} \sqrt{n} F^n(\omega_i^n) \geq \Gamma \right\} = 0.$$

Proof. According to Lemma 3.2 of Dai and He (2010) and the monotonicity of the distribution function $F^n(\cdot)$,

$$\mathbb{P} \left\{ \sup_{0 \leq i \leq E^n(T)} \sqrt{n} F^n(\omega_i^n) \geq \Gamma \right\} \leq \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \sqrt{n} F^n\left(\frac{1}{\sqrt{n}} \tilde{\omega}^n(t)\right) \geq \Gamma \right\}. \quad (2.21)$$

Thus it is enough to prove

$$\lim_{\Gamma \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \sqrt{n} F^n\left(\frac{1}{\sqrt{n}} \tilde{\omega}^n(t)\right) \geq \Gamma \right\} = 0. \quad (2.22)$$

Next we note

$$\mathbb{P} \left\{ \sup_{0 \leq t \leq T} \sqrt{n} F^n\left(\frac{1}{\sqrt{n}} \tilde{\omega}^n(t)\right) \geq \Gamma \right\} \leq \mathbb{P} \left\{ \sqrt{n} F^n\left(\frac{1}{\sqrt{n}} \Gamma_1\right) \geq \Gamma \right\} + \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \tilde{\omega}^n(t) \geq \Gamma_1 \right\}.$$

For any given $\varepsilon > 0$, by Proposition 2.1, we can choose Γ_1 such that

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \tilde{\omega}^n(t) \geq \Gamma_1 \right\} \leq \frac{\varepsilon}{2}.$$

Now from (2.3), for the Γ_1 fixed above,

$$\lim_{\Gamma \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sqrt{n} F^n\left(\frac{1}{\sqrt{n}} \Gamma_1\right) \geq \Gamma \right\} = 0.$$

Thus for any given $\varepsilon > 0$, there is a Γ_0 such that when $\Gamma \geq \Gamma_0$,

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{0 \leq t \leq T} \sqrt{n} F^n\left(\frac{1}{\sqrt{n}} \tilde{\omega}^n(t)\right) \geq \Gamma \right\} \leq \varepsilon. \quad (2.23)$$

This completes the proof of (2.22). Thus the lemma is proved due to (2.21). \square

An immediate consequence of the above lemma is that

$$\mathbb{E} \left[\sup_{0 \leq i \leq E^n(T)} F^n(\omega_i^n) \right] \rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (2.24)$$

This will help to prove Lemma 2.3 below, which is an extension of Proposition 4.2 of Dai and He (2010), where $F^n(\cdot) = F(\cdot)$. The general approach of the proof is the same whether $F^n(\cdot)$'s are the same or vary with n . Roughly speaking, we need to use the martingale convergence theorem (cf. Lemma 4.3 of Dai and He (2010) and Whitt (2007)). The key condition for applying the theorem is (2.24). We thus present the result without repeating the proof.

Lemma 2.3. *Under assumptions (2.1)–(2.5),*

$$\sup_{0 \leq t \leq T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} (\mathbb{1}_{\{\gamma_i^n \leq \omega_i^n\}} - F^n(\omega_i^n)) \cdot h(\omega_i^n) \right| \Rightarrow 0 \text{ as } n \rightarrow \infty,$$

where $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a Borel measurable function such that $0 \leq h(t) \leq 1$ for all $t \in \mathbb{R}_+$.

Similar to Dai et al. (2010), we define the process

$$\zeta^n(t) = \inf\{s \geq 0 : s + \omega^n(s) \geq t\}.$$

It is clear that $\zeta^n \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$ and is nondecreasing for each $n \in \mathbb{Z}_+$.

Lemma 2.4. *Under assumptions (2.1)–(2.5), as $n \rightarrow \infty$*

$$\sup_{0 \leq t \leq T} |\zeta^n(t) - t| \Rightarrow 0.$$

Proof. By the definition of $\zeta^n(t)$, for any $t \geq 0$,

$$0 \leq t - \zeta^n(t) \leq \omega^n(\zeta^n(t)).$$

Hence,

$$\sup_{0 \leq t \leq T} |\zeta^n(t) - t| \leq \sup_{0 \leq t \leq T} \omega^n(\zeta^n(t)) \leq \sup_{0 \leq t \leq T} \omega^n(t).$$

Thus the result follows from Proposition 2.1. \square

Proof of Proposition 2.2. According to Lemma 2.3, it suffices to show that as $n \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \left| \tilde{G}^n(t) - \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} \mathbb{1}_{\{\gamma_i^n \leq \omega_i^n\}} \right| \Rightarrow 0.$$

As a customer arriving at the system before time $\zeta^n(t)$ must have either entered service or abandoned the queue by time t , we have the following relationship:

$$\sum_{i=1}^{E^n(\zeta^n(t)-)} \mathbb{1}_{\{\gamma_i^n \leq \omega_i^n\}} \leq G^n(t) \leq \sum_{i=1}^{E^n(t)} \mathbb{1}_{\{\gamma_i^n \leq \omega_i^n\}}.$$

Hence it is enough to prove that as $n \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \sum_{i=E^n(\zeta^n(t)-)+1}^{E^n(t)} \mathbb{1}_{\{\gamma_i^n \leq \omega_i^n\}} \Rightarrow 0. \quad (2.25)$$

Note that

$$\begin{aligned} \sup_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \sum_{i=E^n(\zeta^n(t)-)+1}^{E^n(t)} \mathbb{1}_{\{\gamma_i^n \leq \omega_i^n\}} &= \sup_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \sum_{i=E^n(\zeta^n(t)-)+1}^{E^n(t)} (\mathbb{1}_{\{\gamma_i^n \leq \omega_i^n\}} - F^n(\omega_i^n)) \\ &\quad + \sup_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \sum_{i=E^n(\zeta^n(t)-)+1}^{E^n(t)} F^n(\omega_i^n). \end{aligned} \quad (2.26)$$

By Lemma 2.3, the first term on the right-hand side of (2.26) will converge to 0. For the second term,

$$\sup_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \sum_{i=E^n(\zeta^n(t)-)+1}^{E^n(t)} F^n(\omega_i^n) \leq \sup_{0 \leq t \leq T} [\bar{E}^n(t) - \bar{E}^n(\zeta^n(t)-)] \cdot \sup_{0 \leq i \leq E^n(T)} \sqrt{n} F^n(\omega_i^n), \quad (2.27)$$

which weakly converges to 0 due to (2.13), Lemmas 2.2 and 2.4. Thus (2.25) holds and the proof is completed. \square

We now present some corollaries which will be used later.

Corollary 2.1. *Under assumptions (2.1)–(2.5), the sequence $\{\tilde{G}^n, n \in \mathbb{Z}_+\}$ is C -tight.*

Proof. According to Proposition 2.2, it is enough to show the C -tightness for $\{\frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} F^n(\omega_i^n), n \in \mathbb{Z}_+\}$. Note that for any $0 \leq s \leq t \leq T$,

$$\frac{1}{\sqrt{n}} \sum_{i=E^n(s)+1}^{E^n(t)} F^n(\omega_i^n) \leq [\bar{E}^n(s) - \bar{E}^n(t)] \cdot \sup_{0 \leq i \leq E^n(T)} \sqrt{n} F^n(\omega_i^n).$$

So the C -tightness follows from the C -tightness of \bar{E}^n (due to (2.13)) and the stochastic boundedness of $\sup_{0 \leq i \leq E^n(T)} \sqrt{n} F^n(\omega_i^n)$ (due to Lemma 2.2). \square

Define the fluid-scaled abandonment process $\bar{G}^n = \{\bar{G}^n(t) : t \geq 0\}$ by

$$\bar{G}^n(t) = \frac{G^n(t)}{n}.$$

Corollary 2.2. *Under assumptions (2.1)–(2.5), as $n \rightarrow \infty$,*

$$\sup_{0 \leq t \leq T} \omega^n(t) \Rightarrow 0 \quad \text{and} \quad \sup_{0 \leq t \leq T} \bar{G}^n(t) \Rightarrow 0.$$

Proof. Note that $\omega^n(t) = \tilde{\omega}^n(t)/\sqrt{n}$ and $\bar{G}^n(t) = \tilde{G}^n(t)/\sqrt{n}$. So the above convergence follows from stochastic boundedness of $\{\tilde{\omega}^n, n \in \mathbb{Z}_+\}$ (Proposition 2.1) and stochastic boundedness of $\{\tilde{G}^n, n \in \mathbb{Z}_+\}$ (Corollary 2.1). \square

Remark 2.2. Dai and He (2010) also prove $\sup_{0 \leq t \leq T} \omega^n(t) \Rightarrow 0$ for the case where $f(x) = \alpha x$, see their Proposition 4.4.

Proof of Proposition 2.3. By the definition of $\omega^n(t)$, we have

$$\begin{aligned} Q^n(t + \omega^n(t)) &\leq E^n(t + \omega^n(t)) - E^n(t) \\ &\leq Q^n((t + \omega^n(t)) -) + \left(E^n(t + \omega^n(t)) - E^n(t + \omega^n(t) - \frac{1}{n}) \right) \\ &\quad + \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} \mathbf{1}_{\{\gamma_i^n \leq \omega_i^n\}}. \end{aligned} \tag{2.28}$$

Note that

$$\frac{1}{\sqrt{n}} (E^n(t + \omega^n(t)) - E^n(t)) = \tilde{E}^n(t + \omega^n(t)) - \tilde{E}^n(t) + \frac{\lambda^n}{n} \cdot \sqrt{n} \cdot \omega^n(t), \tag{2.29}$$

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} \mathbf{1}_{\{\gamma_i^n \leq \omega_i^n\}} &= \frac{1}{\sqrt{n}} \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} \left(\mathbf{1}_{\{\gamma_i^n \leq \omega_i^n\}} - F^n(\omega_i^n) \right) \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} F^n(\omega_i^n). \end{aligned} \tag{2.30}$$

By (2.2) and Corollary 2.2,

$$\sup_{0 \leq t \leq T} |\tilde{E}^n(t + \omega^n(t)) - \tilde{E}^n(t)| \Rightarrow 0, \quad \text{as } n \rightarrow \infty. \tag{2.31}$$

By (2.1) and Proposition 2.1,

$$\left| \frac{\lambda^n}{n} \cdot \sqrt{n} \cdot \omega^n(t) - \mu \tilde{\omega}^n(t) \right| \Rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (2.32)$$

It follows from (2.1) and (2.2) that

$$\begin{aligned} & \sup_{0 \leq t \leq T} \left| E^n(t + \omega^n(t)) - E^n(t + \omega^n(t) - \frac{1}{n}) \right| \\ & \leq \sup_{0 \leq t \leq T} \left| \tilde{E}^n(t + \omega^n(t)) - \tilde{E}^n(t + \omega^n(t) - \frac{1}{n}) \right| + \frac{\lambda^n}{\sqrt{n^3}} \\ & \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned} \quad (2.33)$$

Note that the inequality (2.27) also holds with $(\zeta^n(t)-, t)$ replaced by $(t, t + \omega^n(t))$, so by Corollary 2.2, as $n \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} F^n(\omega_i^n) \Rightarrow 0. \quad (2.34)$$

Lemma 2.3, (2.30) and (2.34) imply that as $n \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \sum_{i=E^n(t)}^{E^n(t+\omega^n(t))} \mathbb{1}_{\{\gamma_i^n \leq \omega_i^n\}} \Rightarrow 0. \quad (2.35)$$

By condition (2.6), as $n \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \left| \tilde{Q}^n(t + \omega^n(t)) - \tilde{Q}^n((t + \omega^n(t)) -) \right| \Rightarrow 0. \quad (2.36)$$

Applying the above convergence (2.31)–(2.36) to the inequality (2.28) yields that as $n \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \left| \tilde{Q}^n(t + \omega^n(t)) - \mu \tilde{\omega}^n(t) \right| \Rightarrow 0.$$

By condition (2.6) and Corollary 2.2, as $n \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \left| \tilde{Q}^n(t + \omega^n(t)) - \tilde{Q}^n(t) \right| \Rightarrow 0.$$

Thus, the result of this proposition follows. \square

3 Diffusion Approximations

We now consider the diffusion approximation for a sequence of $G/GI/N_n + GI$ systems indexed by n . For the n th system, the number of servers is

$$N_n = n^\alpha, \quad (3.1)$$

where $\alpha \in [0, 1]$ is a parameter. Denote $X^n(t)$ as the total number of customers at time t in the n th system. Its diffusion-scaled process $\tilde{X}^n = \{\tilde{X}^n(t) : t \geq 0\}$ is defined by

$$\tilde{X}^n(t) = \frac{X^n(t) - N_n}{\sqrt{n}}. \quad (3.2)$$

It is clear that the queue length process \tilde{Q}^n studied in the previous section satisfies $\tilde{Q}^n = (\tilde{X}^n)^+$. Our main results (Theorem 3.1 (i) and (ii)) characterize the asymptotic behavior of \tilde{X}^n .

In Section 2, we have established an asymptotic relationship between the queue length and abandonment processes with only assumptions on the arrival processes (2.1)–(2.2), patience time distributions (2.3)–(2.4), and the queue length processes (2.5)–(2.6). Some of the intermediate results in Section 2 actually only require (2.1)–(2.5). We will take advantage of the results in Section 2 to obtain the diffusion approximation for the total head count process \tilde{X}^n . To this end, some assumptions on the service time are required. Considering the n th system, let v_i^n , $i = 1, 2, \dots$ be the service time of the i th arriving customer. For $i = -X^n(0) + 1, \dots, -Q^n(0)$, v_i^n denotes the remaining service time of the i th customer initially in service. For $i = -Q^n(0) + 1, \dots, 0$, v_i^n denotes the service time of the i th customer initially waiting in queue. Customer $-Q^n(0) + 1$ is the first in the queue, customer $-Q^n(0) + 2$ is the second, and so on. We assume $\{v_i^n, i \geq -X^n(0) + 1\}$ is a sequence of independent random variables, and is independent of the patience times $\{\gamma_i^n, i \in \mathbb{Z}_+\}$ and the arrival process E^n for each n .

We separate the discussions of service times into two cases. Recall that μ is defined by (2.1). When $0 \leq \alpha < 1$, we assume that the service times follow an exponential distribution.

Assumption 3.1. *The customers' remaining service times and service times $\{v_i^n, i \geq -X^n(0) + 1\}$ are independent and exponentially distributed with rate $\mu^n = n^{1-\alpha}\mu$.*

When $\alpha = 1$, we allow the service times to follow a general distribution function H . Let H_e denote the associated equilibrium distribution, i.e.,

$$H_e(x) = \mu \int_0^x (1 - H(u)) du, \quad x \geq 0,$$

and let M denote the associated renewal function, i.e., M satisfies the following renewal equation

$$M(t) = H(t) + \int_0^t H(t-s) dM(s).$$

Assumption 3.2. *The customer service times $\{v_i^n, i \geq -Q^n(0) + 1\}$ are independent and identically distributed with distribution function $H(\cdot)$ which has mean μ . And the customers' remaining service times $\{v_i^n, -X^n(0) + 1 \leq i \leq -Q^n(0)\}$ are independent and identically distributed with distribution function $H_e(\cdot)$.*

With Assumptions 3.1 and 3.2, the total service capacity of the n th system is $n\mu$, regardless of the choice of α . We assume the following *heavy traffic condition*,

$$\beta^n := \sqrt{n} \left(\frac{\lambda^n}{n\mu} - 1 \right) \rightarrow \beta \quad \text{as } n \rightarrow \infty, \quad (3.3)$$

for some $\beta \in \mathbb{R}$. In particular, the heavy traffic condition implies (2.1). We also assume the convergence of initial states,

$$\tilde{X}^n(0) \Rightarrow \xi \quad \text{as } n \rightarrow \infty, \quad (3.4)$$

for some random variable ξ .

Theorem 3.1. *Assume that conditions (2.2)–(2.4) and (3.3)–(3.4) hold. For the stochastic processes $\{\tilde{X}^n, n \in \mathbb{Z}_+\}$ associated with the sequence of $G/GI/N_n + GI$ systems,*

- (i) *if $0 \leq \alpha < 1$, Assumption 3.1 holds and $\xi \geq 0$ with probability one, then $\tilde{X}^n \Rightarrow \tilde{X}$ as $n \rightarrow \infty$, where \tilde{X} is the solution to the following*

$$\tilde{X}(t) = \xi + \tilde{E}(t) - \sqrt{\mu} \tilde{S}(t) + \beta \mu t - \mu \int_0^t f\left(\frac{(\tilde{X}(s))^+}{\mu}\right) ds + \tilde{L}(t), \quad (3.5)$$

$$\tilde{X}(t) \geq 0, \quad (3.6)$$

$$\tilde{L}(t) \text{ is nondecreasing and } \tilde{L}(0) = 0, \quad (3.7)$$

$$\int_0^\infty \tilde{X}(s) d\tilde{L}(s) = 0, \quad (3.8)$$

and \tilde{S} is a standard Brownian motion independent of \tilde{E} and ξ ;

- (ii) if $\alpha = 1$ and Assumption 3.2 holds, then $\tilde{X}^n \Rightarrow \tilde{X}$ as $n \rightarrow \infty$, where \tilde{X} is the solution to the following

$$\begin{aligned} \tilde{X}(t) = & \xi + \tilde{E}(t) - \tilde{S}(t) + \beta\mu t + \xi^- \cdot (\mu t - M(t)) \\ & + \int_0^t (\tilde{X}(t-s))^- dM(s) - \mu \int_0^t f\left(\frac{(\tilde{X}(s))^+}{\mu}\right) ds, \end{aligned} \quad (3.9)$$

and \tilde{S} is a Gaussian process, which is independent of \tilde{E} and ξ , with the covariance given by

$$\mathbb{E}[\tilde{S}(s)\tilde{S}(t)] = 2 \int_0^s \left(M(u) - u + \frac{1}{2} \right) du + \int_0^s \int_0^t M(s-u)M(t-v)dH(u+v) \quad (3.10)$$

for any $0 \leq s \leq t$.

Remark 3.1. By Proposition 4.9 of Lee and Weerasinghe (2011), we know that for any $y(\cdot) \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$ and a nondecreasing and locally Lipschitz continuous function $g(\cdot)$ defined on \mathbb{R}_+ with $g(0) = 0$, there exist unique $x(\cdot)$ and $\ell(\cdot)$ in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ such that

$$x(t) = y(t) - \int_0^t g((x(s))^+) ds + \ell(t), \quad (3.11)$$

$$x(t) \geq 0, \quad (3.12)$$

$$\ell(\cdot) \text{ is nondecreasing and } \ell(0) = 0, \quad (3.13)$$

$$\text{and } \int_0^\infty x(s) d\ell(s) = 0. \quad (3.14)$$

And by Lemma 5.1, for any $a \in \mathbb{R}$, $y(\cdot) \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$, and a nondecreasing and locally Lipschitz continuous function $g(\cdot)$ defined on \mathbb{R}_+ with $g(0) = 0$, there exists a unique $x(\cdot)$ in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ such that

$$x(t) = a + y(t) + a^- \cdot (\mu t - M(t)) + \int_0^t x(t-s)^- dM(s) - \int_0^t g((x(s))^+) ds.$$

Thus \tilde{X} in Theorem 3.1 (i) and (ii) is well-defined.

4 $G/M/N_n + GI$ Systems with $0 \leq \alpha < 1$

Recall that $G^n(t)$ is the number of customers who have abandoned the queue before time t . The evolution of the process X^n can be characterized by the system dynamics equation

$$X^n(t) = X^n(0) + E^n(t) - S\left(\mu^n \int_0^t (X^n(s) \wedge N_n) ds\right) - G^n(t),$$

where $S(\cdot)$ is a Poisson process with rate one. As a result, we have

$$\begin{aligned} X^n(t) - N_n = & X^n(0) - N_n + E^n(t) - \lambda^n t \\ & - \left[S\left(\mu^n \int_0^t (X^n(s) \wedge N_n) ds\right) - \mu^n \int_0^t (X^n(s) \wedge N_n) ds \right] - G^n(t) \\ & + (\lambda^n - n\mu)t + \mu^n \int_0^t (X^n(s) - N_n)^- ds. \end{aligned} \quad (4.1)$$

Applying the diffusion scaling for X^n, E^n and G^n and the definition of β^n in (3.3), we obtain

$$\tilde{X}^n(t) = \tilde{Y}^n(t) + \mu^n \int_0^t (\tilde{X}^n(s))^- ds - \mu \int_0^t f\left(\frac{1}{\mu}(\tilde{X}^n(s))^+\right) ds, \quad (4.2)$$

where

$$\tilde{Y}^n(t) = \tilde{X}^n(0) + \tilde{E}^n(t) - \tilde{S}^n(t) - \hat{G}^n(t) + \beta^n \mu t, \quad (4.3)$$

with

$$\tilde{S}^n(t) = \frac{1}{\sqrt{n}} \left[S\left(\mu^n \int_0^t (X^n(s) \wedge N_n) ds\right) - \mu^n \int_0^t (X^n(s) \wedge N_n) ds \right], \quad (4.4)$$

$$\hat{G}^n(t) = \tilde{G}^n(t) - \mu \int_0^t f\left(\frac{1}{\mu} \tilde{X}^n(s)^+\right) ds. \quad (4.5)$$

Similarly, let

$$\tilde{Y}^n = \{\tilde{Y}^n(t) : t \geq 0\}, \quad \tilde{S}^n = \{\tilde{S}^n(t) : t \geq 0\}, \quad \hat{G}^n = \{\hat{G}^n(t) : t \geq 0\}.$$

To prove the convergence of process \tilde{X}^n , for any given function $g(\cdot)$ defined on \mathbb{R}_+ , we introduce a sequence of “reflection” mappings $\{\Phi_g^n, n \in \mathbb{Z}_+\}$, where for each n , $\Phi_g^n : \mathbf{D}(\mathbb{R}_+, \mathbb{R}) \rightarrow \mathbf{D}(\mathbb{R}_+, \mathbb{R})$ is defined by $\Phi_g^n(y) = x$ with x being a solution to

$$x(t) = y(t) + \mu^n \int_0^t (x(s))^- ds - \int_0^t g((x(s))^+) ds. \quad (4.6)$$

Note that Φ_g^n depends on n through $\mu^n = n^{1-\alpha}\mu$ based on Assumption 3.1. The following lemma characterizes the sequence of such mappings and allows μ^n to be more general.

Lemma 4.1. (i) *If the function $g(\cdot)$ is a nondecreasing and Lipschitz continuous with $g(0) = 0$, then for each n and $y \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$, there exists a unique solution to (4.6).* (ii) *Let $\{y^n, n \in \mathbb{Z}_+\}$ be a sequence of stochastic processes whose paths lie almost surely in $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$. Assume that $\lim_{n \rightarrow \infty} \mu^n = \infty$ and $\liminf_{n \rightarrow \infty} y^n(0) \geq 0$ with probability one. If the sequence $\{y^n, n \in \mathbb{Z}_+\}$ is C -tight, then the sequence $\{x^n, n \in \mathbb{Z}_+\}$ given by $x^n = \Phi_g^n(y^n)$ is also C -tight, and any limit (x^*, y^*) of a converging subsequence of $\{(x^n, y^n), n \in \mathbb{Z}_+\}$ satisfies (iia) $x^*(t) \geq 0$; and (iib) If $\ell^*(t) = x^*(t) + \int_0^t g(x^*(s)^+) ds - y^*(t)$, in other words, ℓ^* is the limit of a converging subsequence of $\{\mu^n \int_0^t x^n(s)^- ds, n \in \mathbb{Z}_+\}$, then $\int_0^t x^*(s) d\ell^*(s) = 0$.*

Proof. First we define a sequence of mappings $\{\Psi_g^n, n \in \mathbb{Z}_+\}$ with $\Psi_g^n : \mathbf{D}(\mathbb{R}_+, \mathbb{R}) \rightarrow \mathbf{D}(\mathbb{R}_+, \mathbb{R})$ by $\Psi_g^n(x)(t) = \mu^n x^-(t) - g(x^+(t))$ for each $x \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$. Clearly, Ψ_g^n is Lipschitz continuous for each n since $g(0) = 0$ and $g(\cdot)$ is Lipschitz continuous. Therefore, by Lemma 1 of Reed and Ward (2004), there exists a unique solution to (4.6). Thus we have (i).

Now we prove (ii). Note that by Lemma 1 of Reed and Ward (2004), if the input process $y^n \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$, then the resulting process $x^n \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$. We define z^n to be the solution to

$$z^n(t) = y^n(t) - \mu^n \int_0^t z^n(s) ds. \quad (4.7)$$

Again, Lemma 1 of Reed and Ward (2004) guarantees the existence and uniqueness of z^n . Since $g(\cdot)$ is Lipschitz continuous, there exists n_0 such that $g(x) \leq \mu^n x$ for all $x \geq 0$ and $n > n_0$. Since

$$x^n(t) - z^n(t) = \mu^n \int_0^t (x^n(s))^- ds - \int_0^t g((x^n(s))^+) ds + \mu^n \int_0^t z^n(s) ds,$$

$x^n(\cdot) - z^n(\cdot)$ is differentiable and

$$\begin{aligned} \frac{d}{dt}[x^n(t) - z^n(t)] &= -g(x^n(t)^+) + \mu^n [x^n(t)^- + z^n(t)] \\ &\geq -\mu^n [x^n(t) - z^n(t)]. \end{aligned}$$

As a result, $\frac{d}{dt}[(x^n(t) - z^n(t)) \exp(\mu^n t)] \geq 0$. Since $(x^n(0) - z^n(0)) \exp(\mu^n 0) = 0$, we have $(x^n(t) - z^n(t)) \exp(\mu^n t) \geq 0$. This implies that

$$x^n(t) \geq z^n(t) \quad \text{for all } t \geq 0 \text{ and } n > n_0. \quad (4.8)$$

The integral equation (4.7) can easily be solved to obtain

$$z^n(t) = \exp(-\mu^n t) \left[y^n(0) + \int_0^t \exp(\mu^n s) dy^n(s) \right].$$

Performing integration by parts yields

$$y^n(0) + \int_0^t \exp(\mu^n s) dy^n(s) = y^n(t) + \mu^n \int_0^t [y^n(t) - y^n(s)] \exp(\mu^n s) ds.$$

This implies that for any $0 < \delta < t$,

$$\begin{aligned} z^n(t) &= y^n(t) \exp(-\mu^n t) + \mu^n \exp(-\mu^n t) \int_0^t (y^n(t) - y^n(s)) \exp(\mu^n s) ds \\ &\geq y^n(t) \exp(-\mu^n t) - \sup_{t-\delta \leq s \leq t} |y^n(t) - y^n(s)| - 2 \exp(-\mu^n \delta) \cdot \sup_{0 \leq s \leq t-\delta} |y^n(s)|. \end{aligned}$$

Hence, for any $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P} \left\{ \sup_{t \in [0, T]} (z^n(t))^- \geq \varepsilon \right\} &\leq \mathbb{P} \left\{ \sup_{t \in [0, \delta]} (y^n(t))^- > \frac{\varepsilon}{3} \right\} \\ &\quad + \mathbb{P} \left\{ \sup_{t \in [\delta, T]} (y^n(t))^- \cdot \exp(-\mu^n t) > \frac{\varepsilon}{3} \right\} \\ &\quad + \mathbb{P} \left\{ \sup_{t \in [0, T]} \sup_{t-\delta \leq s \leq t} |y^n(t) - y^n(s)| > \frac{\varepsilon}{3} \right\} \\ &\quad + \mathbb{P} \left\{ \sup_{t \in [0, T]} 2 \exp(-\mu^n \delta) \cdot \sup_{0 \leq s \leq t-\delta} |y^n(s)| > \frac{\varepsilon}{3} \right\}. \end{aligned} \quad (4.9)$$

Note that the C -tightness of $\{y^n, n \in \mathbb{Z}_+\}$ implies the C -tightness of $\{(y^n)^-, n \in \mathbb{Z}_+\}$, and $\liminf_{n \rightarrow \infty} y^n(0) \geq 0$ implies that $\lim_{n \rightarrow \infty} (y^n(0))^- = 0$. Hence, we have

$$\lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{t \in [0, \delta]} (y^n(t))^- > \frac{\varepsilon}{3} \right\} = 0. \quad (4.10)$$

Similarly, the C -tightness of $\{y^n, n \in \mathbb{Z}_+\}$ and $\{(y^n)^-, n \in \mathbb{Z}_+\}$, and $\lim_{n \rightarrow \infty} \mu^n = \infty$ make the last three terms in (4.9) converge to zero. Therefore, by (4.10),

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{t \in [0, T]} (z^n(t))^- \geq \varepsilon \right\} = 0.$$

It follows from (4.8) that

$$\sup_{t \in [0, T]} (x^n(t))^- \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (4.11)$$

Define

$$\Phi_g(y)(t) = x(t) \quad \text{and} \quad \Psi_g(y)(t) = \ell(t), \quad (4.12)$$

where $x(\cdot)$ and $\ell(\cdot)$ are given by (3.11)–(3.14) in Remark 3.1. Now note that

$$(x^n(t))^+ = y^n(t) + (x^n(s))^- - \int_0^t g((x^n(s))^+) ds + \mu^n \int_0^t (x^n(s))^- ds.$$

Hence,

$$(x^n)^+ = \Phi_g(y^n + (x^n)^-) \quad \text{and} \quad \mu^n \int_0^\cdot (x^n(s))^- ds = \Psi_g(y^n + (x^n)^-)(\cdot).$$

In view of (4.11), we have that

$$\{y^n + (x^n)^-, n \in \mathbb{Z}_+\} \text{ is } C\text{-tight.} \quad (4.13)$$

By Proposition 4.9 of Lee and Weerasinghe (2011), the mappings Φ_g and Ψ_g are Lipschitz continuous. By the continuous mapping theorem, (4.13) implies that $\{(x^n)^+, n \in \mathbb{Z}_+\}$ is C -tight, hence $\{x^n, n \in \mathbb{Z}_+\}$ is C -tight by (4.11).

(iia) follows directly from (4.11). Suppose (iib) does not hold, then there exists a point t such that $x^*(t) = \varepsilon > 0$ and $\ell^*(\cdot)$ is increasing at t . This implies that there exists a $\delta > 0$ such that for all large enough n , $\ell^n(t + \delta) > \ell^n(t - \delta)$ and $x^n(s) > \varepsilon/2$ for all $s \in (t - \delta, t + \delta)$. On the other hand, however, when $x^n(s) > \varepsilon/2$ for all $s \in (t - \delta, t + \delta)$, we have that $(x^n)^-(s) = 0$ for all $s \in (t - \delta, t + \delta)$. This consequently implies $\ell^n(t - \delta) = \ell^n(t + \delta)$. So we get a contradiction. Hence, (iib) holds. \square

In view of (4.2) and (4.6), it is clear that $\tilde{X}^n = \Phi_g^n(\tilde{Y}^n)$. A key step in applying Lemma 4.1 to obtain the diffusion limit is to show the C -tightness of $\{\tilde{Y}^n, n \in \mathbb{Z}_+\}$. The involvement of the abandonment process in \tilde{Y}^n makes the task challenging. To this end, we first present a study on the systems without abandonment. The result is not only of independent interest, but also paves the way for studying the systems with abandonment.

4.1 Systems without Abandonment

In this section, we consider the $G/M/N_n$ system. The arrival process to the n th system is still E^n . Customers will not leave the queue until they have obtained service, so the abandonment process is equal to 0. To differentiate, let $X_0^n(t)$ be the total number of customers at time t in the n th system without abandonment. Similarly, the process $\tilde{X}_0^n = \{\tilde{X}_0^n(t) : t \geq 0\}$ is defined as $\tilde{X}_0^n(t) = (X_0^n(t) - N_n)/\sqrt{n}$. According to our convention given in Section 2 (see the paragraph between (2.2) and (2.3)), we have

$$\tilde{X}_0^n(0) = \tilde{X}^n(0).$$

The objective of this section is to establish the weak convergence of $\{\tilde{X}_0^n, n \in \mathbb{Z}_+\}$ (see Proposition 4.2). Let $\tilde{S}_0^n = \{\tilde{S}_0^n(t) : t \geq 0\}$ with

$$\tilde{S}_0^n(t) = \frac{1}{\sqrt{n}} \left[S\left(\mu^n \int_0^t (X_0^n(s) \wedge N_n) ds\right) - \mu^n \int_0^t (X_0^n(s) \wedge N_n) ds \right].$$

In such systems, the diffusion-scaled balance equation (4.2) becomes

$$\tilde{X}_0^n(t) = \tilde{Y}_0^n(t) + \mu^n \int_0^t (\tilde{X}_0^n(s))^- ds, \quad (4.14)$$

where

$$\tilde{Y}_0^n(t) = \tilde{X}_0^n(0) + \tilde{E}^n(t) - \tilde{S}_0^n(t) + \beta^n \mu t.$$

We specialize the function $g \equiv 0$ in Lemma 4.1 and denote by Φ_0^n the corresponding reflection mapping for the n th system. So we have $\tilde{X}_0^n = \Phi_0^n(\tilde{Y}_0^n)$.

Proposition 4.1. *Assume that conditions (2.2), (3.3)-(3.4), and Assumption 3.1 hold. For the sequence of $G/M/N_n$ systems with $0 \leq \alpha < 1$, as $n \rightarrow \infty$,*

$$\tilde{S}_0^n \Rightarrow \sqrt{\mu} \tilde{S}_0,$$

where \tilde{S}_0 is a standard Brownian motion independent of the initial state and the arrival process.

Proof. By (3.1) and Assumption 3.1,

$$\mu^n \int_0^t (X_0^n(s) \wedge N_n) ds \leq nt. \quad (4.15)$$

Thus

$$\mathbb{P} \left\{ \sup_{t \in [0, T]} |\tilde{S}_0^n(t)| > \Gamma \right\} \leq \mathbb{P} \left\{ \sup_{t \in [0, T]} \frac{1}{\sqrt{n}} |S(nt) - nt| > \Gamma \right\}, \quad (4.16)$$

which vanishes as $\Gamma \rightarrow \infty$ by the central limit theorem for Poisson processes (a special case of Theorem 17.3 of Billingsley (1999)). So $\{\tilde{S}_0^n, n \in \mathbb{Z}_+\}$ is stochastically bounded. Other components of $\{\tilde{Y}_0^n, n \in \mathbb{Z}_+\}$ are stochastically bounded because of the assumptions (2.2), (3.3)-(3.4). Noting that

$$\sup_{t \in [0, T]} (\tilde{X}_0^n(t))^- \leq \sup_{t \in [0, T]} (\tilde{Y}_0^n(t))^-,$$

we have stochastic boundedness for $\{\sup_{t \in [0, T]} (\tilde{X}_0^n(t))^-, n \in \mathbb{Z}_+\}$. Furthermore, by

$$(\tilde{X}_0^n(t))^+ = \tilde{Y}_0^n(t) + (\tilde{X}_0^n(t))^- + \mu^n \int_0^t (\tilde{X}_0^n(s))^- ds,$$

we have $(\tilde{X}_0^n)^+ = \Phi_g(\tilde{Y}_0^n + (\tilde{X}_0^n)^-)$ with $g \equiv 0$ (see (4.12)). From the Lipschitz continuity of the mapping $\Phi_g(\cdot)$ (see Proposition 2 of Reed and Ward (2004)), we have stochastic boundedness for $\{(\tilde{X}_0^n)^+, n \in \mathbb{Z}_+\}$. In view of $\tilde{X}_0^n = (\tilde{X}_0^n)^+ - (\tilde{X}_0^n)^-$, therefore, $\{\tilde{X}_0^n, n \in \mathbb{Z}_+\}$ is also stochastically bounded. The stochastic boundedness immediately implies that

$$\bar{X}_0^n \Rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad (4.17)$$

where $\bar{X}_0^n = \{\frac{X_0^n(t) - N_n}{n} : t \geq 0\}$. We now prove that

$$\int_0^\cdot \mu^n (\bar{X}_0^n(s))^- ds \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (4.18)$$

It follows from (4.14) that

$$\begin{aligned} \int_0^t \mu^n (\bar{X}_0^n(s))^- ds &= \bar{X}_0^n(t) - \bar{X}_0^n(0) - \frac{E^n(t) - \lambda^n t}{n} - \frac{(\lambda^n - n\mu)t}{n} \\ &\quad + \frac{S\left(\mu^n \int_0^t (X_0^n(s) \wedge N_n) ds\right) - \mu^n \int_0^t (X_0^n(s) \wedge N_n) ds}{n}. \end{aligned} \quad (4.19)$$

By (4.17), (2.2) and (3.3), the first four terms on the right hand side of the above converge to zero in distribution. By (4.15)–(4.16), the last term in (4.19) also converges to 0 in probability.

By $(X_0^n \wedge N_n)/n = N_n/n - (\bar{X}_0^n)^-$, (4.18) is equivalent to

$$\mu^n \int_0^\cdot \frac{X_0^n(s) \wedge N_n}{n} ds \Rightarrow \bar{e}(\cdot),$$

where $\bar{e}(\cdot)$ is given by (2.13), the proposition follows from the FCLT for renewal processes (cf. Theorem 17.3 of Billingsley (1999)) and the random-time-change theorem (cf. Corollary 1 of Whitt (1980)). \square

Proposition 4.2. *Assume that conditions (2.2), (3.3)-(3.4), and Assumption 3.1 hold. For the sequence of $G/M/N_n$ systems with $0 \leq \alpha < 1$,*

$$\tilde{X}_0^n \Rightarrow \tilde{X}_0 \quad \text{as } n \rightarrow \infty,$$

where \tilde{X}_0 is the solution to the following

$$\tilde{X}_0(t) = \xi + \tilde{E}(t) - \tilde{S}_0(\mu t) + \beta \mu t + \tilde{L}_0(t), \quad (4.20)$$

$$\tilde{X}_0(t) \geq 0, \quad (4.21)$$

$$\tilde{L}_0 \text{ is nondecreasing and } \tilde{L}_0(0) = 0, \quad (4.22)$$

$$\int_0^\infty \tilde{X}_0(s) d\tilde{L}_0(s) = 0, \quad (4.23)$$

where \tilde{S}_0 is given by Proposition 4.1 and is independent of \tilde{E} and ξ .

Proof. Conditions (2.2), (3.3)-(3.4), and Proposition 4.1 imply that $\{\tilde{Y}_0^n, n \in \mathbb{Z}_+\}$ is C -tight and as $n \rightarrow \infty$,

$$\tilde{Y}_0^n \Rightarrow \xi + \tilde{E} - \sqrt{\mu} \tilde{S}_0 + \beta \bar{e}, \quad (4.24)$$

where $\bar{e}(\cdot)$ is given by (2.13). By the proof of Lemma 4.1 with $g = 0$ (see (4.11) and (4.12)), we know that

$$(\tilde{X}_0^n)^- \Rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad (4.25)$$

and

$$(\tilde{X}_0^n)^+ = \Phi_0(\tilde{Y}_0^n + (\tilde{X}_0^n)^-) \quad \text{and} \quad \mu^n \int_0^t (\tilde{X}_0^n(s))^- ds = \Psi_0(\tilde{Y}_0^n + (\tilde{X}_0^n)^-)(t).$$

The theorem directly follows from the Lipschitz continuity of Φ_0 and Ψ_0 , the random-time-change theorem (cf. Corollary 1 of Whitt (1980)), and (4.24)–(4.25). \square

4.2 Systems with Abandonment

We now return to the study of systems with abandonment. Proposition 4.2 not only remains as an independent result, but also helps to establish a useful bound for studying systems with abandonment. Recall that $\tilde{Q}^n = (\tilde{X}^n)^+$ and $\tilde{Q}_0^n = (\tilde{X}_0^n)^+$. By Theorem 2.2 of Dai and He (2010), on each sample path,

$$\tilde{Q}^n(t) \leq \tilde{Q}_0^n(t) \quad \text{for all } t \geq 0. \quad (4.26)$$

Since Proposition 4.2 has already shown the stochastic boundedness of $\{\tilde{Q}_0^n, n \in \mathbb{Z}_+\}$, by (4.26), $\{\tilde{Q}^n, n \in \mathbb{Z}_+\}$ is stochastically bounded.

Similar to the study for systems without abandonment, the stochastic boundedness of the queue length processes helps to prove the following convergence for the diffusion-scaled service process \tilde{S}^n defined in (4.4).

Proposition 4.3. *Assume that conditions (2.2), (3.3)-(3.4), and Assumption 3.1 hold. For the sequence of $G/M/N_n + GI$ systems with $0 \leq \alpha < 1$, as $n \rightarrow \infty$,*

$$\tilde{S}^n \Rightarrow \sqrt{\mu} \tilde{S}, \quad (4.27)$$

where \tilde{S} is a standard Brownian motion independent of the initial state and the arrival process.

Proof. The proof is essentially the same as that for Proposition 4.1, so we only point out the differences.

Since $\{\tilde{Q}^n, n \in \mathbb{Z}_+\}$ is stochastically bounded, we immediately know that $(X^n(\cdot) - N_n)^+/n \Rightarrow 0$ as $n \rightarrow \infty$. Thus, in view of the fact that $(X^n(\cdot) - N_n)^-/n \leq N_n/n \rightarrow 0$ as $n \rightarrow \infty$, we have

$$\frac{X^n(\cdot) - N_n}{n} \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (4.28)$$

Next we prove

$$\int_0^t \frac{\mu^n(X^n(s) - N_n)^-}{n} ds \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (4.29)$$

It follows from (4.1) that

$$\begin{aligned} \int_0^t \frac{\mu^n(X^n(s) - N_n)^-}{n} ds &= \frac{X^n(t) - N_n}{n} - \frac{X^n(0) - N_n}{n} - \frac{E^n(t) - \lambda^n t}{n} - \frac{(\lambda^n - n\mu)t}{n} \\ &\quad + \frac{S\left(\mu^n \int_0^t X^n(s) \wedge N_n ds\right) - \mu^n \int_0^t X^n(s) \wedge N_n ds}{n} \\ &\quad + \frac{G^n(t)}{n}. \end{aligned} \quad (4.30)$$

The only difference between (4.30) and (4.19) is the term $G^n(t)/n$, which converges to zero in distribution by Corollary 2.2. So (4.29) holds by the same argument as (4.18). The same application of the FCLT and the random-time-change theorem then yields (4.27). \square

Proof of Theorem 3.1 (i). The sequence of the queue length processes is stochastically bounded by (4.26) and Proposition 4.2. Since $f(\cdot)$ is continuous, the sequence $\{\mu \int_0^t f((\tilde{X}^n(s))^+/\mu) ds, n \in \mathbb{Z}_+\}$ (see (4.5)) is C -tight. Then the C -tightness of the sequence $\{\tilde{Y}^n, n \in \mathbb{Z}_+\}$ follows from condition (3.4) on the initial states, condition (2.2) on the arrival process, Proposition 4.3 for the service process, Corollary 2.1 for the abandonment process, and the heavy traffic condition (3.3). Note that the above reasoning relies only on stochastic boundedness of the queue length process, not on the modulus of continuity being asymptotically small (see Figure 1.1).

Applying Lemma 4.1 we have the C -tightness of $\{\tilde{X}^n, n \in \mathbb{Z}_+\}$. This implies the C -tightness of $\{\tilde{Q}^n, n \in \mathbb{Z}_+\}$. By Theorem 2.1, we have that $\hat{G}^n(t) \Rightarrow 0$ as $n \rightarrow \infty$. This implies that as $n \rightarrow \infty$,

$$\tilde{Y}^n \Rightarrow \xi + \tilde{E} - \sqrt{\mu} \tilde{S} + \beta \bar{e}. \quad (4.31)$$

By the proof of Lemma 4.1 with $g(\cdot) = \mu f(\cdot/\mu)$ (see (4.11) and (4.12)), we know that

$$(\tilde{X}^n)^- \Rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad (4.32)$$

and

$$\tilde{X}^n = \Phi_g(\tilde{Y}^n + (\tilde{X}^n)^-) \quad \text{and} \quad \mu^n \int_0^t (\tilde{X}^n(s))^- ds = \Psi_g(\tilde{Y}^n + (\tilde{X}^n)^-)(t).$$

The theorem directly follows from the Lipschitz continuity of Φ_g and Ψ_g , the random-time-change theorem (cf. Corollary 1 of Whitt (1980)), and (4.31)–(4.32). \square

5 $G/GI/N_n + GI$ Systems with $\alpha = 1$

Now we consider the case where $\alpha = 1$, i.e., the number of servers $N_n = n$. Denote $S^n(t)$ as the number of customers leaving the n th system by finishing their services. We then have the following simple balance equation

$$X^n(t) = X^n(0) + E^n(t) - S^n(t) - G^n(t). \quad (5.1)$$

The asymptotic behavior of E^n under diffusion scaling follows directly from the assumption on the arrival processes. In what follows, we obtain the asymptotic behavior of X^n by studying that of S^n and G^n under diffusion scaling.

The asymptotic analysis of S^n follows from the ideas of Reed (2007), with some technical results from Reed (2009) and Krichagina and Puhalskii (1997). We first present a set of relationships from Reed (2007) which is applicable for all $G/GI/N_n + G$ systems. Recently, the same set of relationships is also applied to study a multi-class model in Reed and Shaki (2012). Let $K^n(t)$ be the number of customers who have entered service by time t . Denote by κ_i^n the i th jump time of the counting process K^n . Define

$$M^n(t) = \sum_{i=-Q^n(0)+1}^{K^n(t)-Q^n(0)} \left(\mathbb{1}_{\{\kappa_i^n + v_i^n > t\}} - H^c(t - \kappa_i^n) \right). \quad (5.2)$$

and

$$W_0^n(t) = \sum_{j=-X^n(0)+1}^{-Q^n(0)} \left(\mathbb{1}_{\{v_j^n > t\}} - H_e^c(t) \right). \quad (5.3)$$

According to Propositions 2, 5 and 6 of Reed (2007), the service completion process is

$$S^n(t) = \hat{S}^n(t) + (X_0^n \wedge n) \cdot \mu t + (X_0^n - n)^- \cdot M(t) - \int_0^t (X^n(t-s) - n)^- dM(s), \quad (5.4)$$

where

$$\hat{S}^n(t) = - (W_0^n(t) + M^n(t)) - \int_0^t (W_0^n(t-s) + M^n(t-s)) dM(s). \quad (5.5)$$

This relationship was initially obtained in Sections 2 and 3 of Reed (2007) in the context of a multi-server queue without abandonment by manipulating system dynamic equations in an ingenious manner, without involving any asymptotic analysis. It is, however, still true for queues with abandonment, since what (5.4) and (5.5) reveal is a relationship between how customers enter service – the process K^n , and how they complete service – the process S^n . The process K^n plays the role of separating the buffer and the server pool. Of course, due to the differences between models with and without abandonment, the asymptotic study of K^n may need different methods. Plugging (5.4) into (5.1) yields

$$\begin{aligned} X^n(t) &= X^n(0) + E^n(t) - G^n(t) - \hat{S}^n(t) \\ &\quad - (X_0^n \wedge n) \mu t - (X_0^n - n)^- \cdot M(t) + \int_0^t (X^n(t-s) - n)^- dM(s). \end{aligned} \quad (5.6)$$

Thus (5.6) implies that

$$\tilde{X}^n(t) = \tilde{Y}^n(t) + \int_0^t (\tilde{X}^n(t-s))^- dM(s) - \mu \int_0^t f\left(\frac{1}{\mu}(\tilde{X}^n(s))^+\right) ds, \quad (5.7)$$

where \hat{G}^n is defined as in (4.5) and

$$\tilde{Y}^n(t) = \tilde{X}^n(0) + \tilde{E}^n(t) - \hat{G}^n(t) - \tilde{S}^n(t) + \beta^n \mu t + \tilde{X}^n(0)^- \cdot (\mu t - M(t)), \quad (5.8)$$

$$\tilde{S}^n(t) = \frac{\hat{S}^n(t)}{\sqrt{n}}. \quad (5.9)$$

To further study the process \tilde{X}^n , we introduce a reflection mapping in the following lemma.

Lemma 5.1. *Assume that g is a locally Lipschitz continuous function with $g(0) = 0$. For any $y \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$, there exists a unique solution x to the following equation*

$$x(t) = y(t) + \int_0^t (x(t-s))^- dM(s) + \int_0^t g((x(s))^+) ds. \quad (5.10)$$

Moreover, the mapping $\Phi_{M,g} : \mathbf{D}(\mathbb{R}_+, \mathbb{R}) \rightarrow \mathbf{D}(\mathbb{R}_+, \mathbb{R})$ defined by $x = \Phi_{M,g}(y)$ is Lipschitz continuous in the topology of uniform convergence over bounded intervals, measurable with respect to the Borel σ -field generated by the Skorohod J_1 -topology.

The proof of this lemma will be presented in Appendix A. Note that the mapping $\Phi_{M,g}$ is an extension of the one in Section 4 of Reed (2007). Following this lemma, we have $\tilde{X}^n = \Phi_{M,g}(\tilde{Y}^n)$ with $g(t) = \mu f(t/\mu)$. The next proposition treats the convergence of diffusion-scaled service processes jointly with the arrival process and initial status.

Proposition 5.1. *Assume that conditions (2.2), (3.3)–(3.4) and Assumption 3.2 hold. For the sequence of $G/GI/n + GI$ systems,*

$$(\tilde{X}^n(0), \tilde{E}^n, \tilde{S}^n) \Rightarrow (\xi, \tilde{E}, \tilde{S}), \quad (5.11)$$

where ξ , \tilde{E} and \tilde{S} are independent of each other, and \tilde{S} is a Gaussian process with continuous sample paths, zero mean and covariance function given by

$$\mathbb{E}\tilde{S}(t)\tilde{S}(t+\delta) = 2 \int_0^t (M(u) - u + \frac{1}{2}) du + \int_0^t \int_0^{t+\delta} M(t-u)M(t+\delta-v) dF(u+v) \quad \text{for } t, \delta \geq 0.$$

This result is essentially the same as Proposition 9 of Reed (2007), and in fact the proofs for the two are almost identical. What prevents us from directly citing the result is the definition of the term $M^n(t)$, which involves the process $K^n(t)$ of customers entering service. For a multi-server queue with or without abandonment, the process of customers entering service will be different. Let the processes $\tilde{M}_n = \{\tilde{M}_n(t) : t \geq 0\}$ and $\tilde{K}^n = \{\tilde{K}^n(t) : t \geq 0\}$ be defined by $\tilde{M}^n(t) = M^n(t)/\sqrt{n}$ and $\tilde{K}^n(t) = K^n(t)/n$, respectively. According to the separation of variation pointed out in Krichagina and Puhalskii (1997), the diffusion limit of $\{\tilde{M}^n, n \in \mathbb{Z}_+\}$ is only affected by the fluid limit of $\{\tilde{K}^n, n \in \mathbb{Z}_+\}$. It turns out that, according to the following lemma, the fluid limit of $\{\tilde{K}^n, n \in \mathbb{Z}_+\}$ is the same as that for queues without abandonment in the Halfin-Whitt regime.

Lemma 5.2. *Under conditions (2.2) and (3.3)–(3.4), and Assumption 3.2, we have that*

$$\tilde{K}^n \Rightarrow \bar{e} \quad \text{as } n \rightarrow \infty, \quad (5.12)$$

where \bar{e} given by (2.13).

Proof. The limit of $\tilde{X}^n(0)$ being stochastically bounded implies that $(X^n(0) - n)/n \Rightarrow 0$ as $n \rightarrow \infty$. By Assumption 3.2, according to Theorems 3.2 and 3.3 of Zhang (2012), $\tilde{K}^n \Rightarrow \bar{e}$ as $n \rightarrow \infty$. \square

To facilitate explaining why only the fluid limit of K^n determines the diffusion limit of S^n , we briefly sketch the reasoning that originated from Krichagina and Puhalskii (1997). According

to (5.2),

$$\begin{aligned}
\tilde{M}^n(t) &= \frac{1}{\sqrt{n}} \sum_{i=-Q^n(0)+1}^{K^n(t)-Q^n(0)} \left(\mathbb{1}_{\{\kappa_i^n + v_i^n > t\}} - H^c(t - \kappa_i^n) \right) \\
&= -\frac{1}{\sqrt{n}} \sum_{i=-Q^n(0)+1}^{K^n(t)-Q^n(0)} \left(\mathbb{1}_{\{\kappa_i^n + v_i^n \leq t\}} - H(t - \kappa_i^n) \right) \\
&= -\int_0^t \int_0^t \mathbb{1}_{\{s+x \leq t\}} dV^n(s, x),
\end{aligned}$$

where

$$V^n(s, x) = \frac{1}{\sqrt{n}} \sum_{i=-Q^n(0)+1}^{K^n(s)-Q^n(0)} \left(\mathbb{1}_{\{v_i^n \leq x\}} - H(x) \right). \quad (5.13)$$

Here $-\tilde{M}^n$ is defined in the same way as in Krichagina and Puhalskii (1997) using stochastic integral, and the fluid limit of $\tilde{K}^n(\cdot)$ plays the same role as $a(\cdot)$ in Lemma 5.3 of Krichagina and Puhalskii (1997). We thus have the following diffusion limit of $-\tilde{M}^n$, which is cited from Lemma 5.3 and Remark 4 of Krichagina and Puhalskii (1997) without repeating the proofs.

Lemma 5.3. *If (5.12) holds, then*

$$-\tilde{M}^n \Rightarrow \tilde{M} \quad \text{as } n \rightarrow \infty, \quad (5.14)$$

where \tilde{M} is a Gaussian process with continuous sample paths, zero mean and covariance function given by

$$\mathbb{E} \tilde{M}(t) \tilde{M}(t + \delta) = \int_0^t H(t - u) [1 - H(t + \delta - u)] \mu du. \quad (5.15)$$

We can now prove Proposition 5.1 by connecting it to Proposition 9 of Reed (2007) and some of the results of Reed (2009).

Proof of Proposition 5.1. Based on Lemmas 5.2 and 5.3, we have $\tilde{M}^n \Rightarrow \tilde{M}$ as $n \rightarrow \infty$. The negative sign in (5.14) does not matter since \tilde{M}_2 is Gaussian with zero mean. Now, define

$$\hat{M}^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor \mu n t \rfloor} \left(\mathbb{1}_{\{\frac{k}{n\mu} + v_i^n > t\}} - H^c(t - \frac{k}{\mu n}) \right).$$

According to Lemma 5.2 and the argument in Proposition 5.1 of Reed (2009), we have joint convergence $(\hat{M}^n, \tilde{M}^n) \Rightarrow (\tilde{M}, \tilde{M})$ as $n \rightarrow \infty$. This implies Proposition 5.2 of Reed (2009), which gives the joint convergence of

$$(\tilde{X}^n(0), \tilde{E}^n, \tilde{W}^n, \tilde{M}^n) \Rightarrow (\xi, \tilde{E}, \tilde{W}(H_e), \tilde{M}) \quad \text{as } n \rightarrow \infty.$$

Note that here our \tilde{W}^n and $\tilde{W}(H_e)$ are the same as those of Reed (2007, 2009), since these two terms depend only on the initial status of the server pool. Thus, exactly the same approach from Proposition 9 of Reed (2007) applies here. Roughly speaking, the idea is to apply the continuous mapping to map $(\tilde{X}^n(0), \tilde{E}^n, \tilde{W}^n, \tilde{M}^n)$ to $(\tilde{X}^n(0), \tilde{E}^n, \tilde{S}^n)$. We omit the repeat of the same argument. \square

Let \tilde{Q}_0^n denote the queue length process of the many-server system without abandonment. It is proved in Reed (2009) that $\{\tilde{Q}_0^n, n \in \mathbb{Z}_+\}$ is stochastically bounded. Again, by Theorem 2.2 of Dai and He (2010), with probability one, $\tilde{Q}^n(t) \leq \tilde{Q}_0^n(t)$ for all $t \geq 0$. This implies that

$\{\tilde{Q}^n, n \in \mathbb{Z}_+\}$ is stochastically bounded. We now can apply the similar argument for the system with $\alpha < 1$ to the system with $\alpha = 1$. The basic logic, which was depicted in Figure 1.1, is now further elaborated as follows. With the help of Corollary 2.1 the stochastic boundedness of the queue length process implies the C -tightness of the centered abandonment process $\{\hat{G}^n, n \in \mathbb{Z}_+\}$. This, together with the standard assumptions, shows that the input process $\{\tilde{Y}^n, n \in \mathbb{Z}_+\}$ of the mapping $\Phi_{M,g}$ is C -tight. Utilizing the nice properties of the regulator map $\Phi_{M,g}$, we obtain the convergence of the output process $\{\tilde{X}^n, n \in \mathbb{Z}_+\}$. In the following, we give the rigorous proof.

Proof of Theorem 3.1 (ii). Since the sequence of the queue length processes is stochastically bounded and $f(\cdot)$ is continuous, $\{\mu \int_0^t f((\tilde{X}^n(s))^+/\mu) ds, n \in \mathbb{Z}_+\}$ is C -tight. By Corollary 2.1, the abandonment process $\{\tilde{G}^n, n \in \mathbb{Z}_+\}$ is also C -tight. According to the definition of (4.5), $\{\hat{G}^n, n \in \mathbb{Z}_+\}$ is C -tight. Then the C -tightness of the processes $\{\tilde{Y}^n, n \in \mathbb{Z}_+\}$ follows from condition (3.4) on the initial states, condition (2.2) on the arrival processes, Proposition 5.1 for the service processes, and the heavy traffic condition (3.3).

In order to prove the convergence of $\{\tilde{X}^n, n \in \mathbb{Z}_+\}$ and that the corresponding limit is given by (3.9), by the uniqueness of the solution to the mapping $\Phi_{M,g}$ in Lemma 5.1, it is enough to show that for every subsequence $\{\tilde{X}^{n_k}, n_k \in \mathbb{Z}_+\}$, there exists a further subsequence converging to $\Phi_{M,g}(\tilde{Y})$ with

$$\tilde{Y} = \{\tilde{Y}(t) : t \geq 0\} = \{\xi + \tilde{E}(t) - \tilde{S}(t) + \beta\mu t + \xi^- \cdot (\mu t - M(t)) : t \geq 0\}.$$

Since the sequence $\{\tilde{Y}^n, n \in \mathbb{Z}_+\}$ is C -tight, for the subsequence $\{\tilde{Y}^{n_k}, n_k \in \mathbb{Z}_+\}$, there is a further convergent subsequence $\{\tilde{Y}^{n'_k}, n'_k \in \mathbb{Z}_+\}$, along which

$$\tilde{Y}^{n'_k} \Rightarrow \tilde{Y}^\infty \quad \text{as } n'_k \rightarrow \infty, \quad (5.16)$$

in the Skorohod J_1 -topology for some limit $\tilde{Y}^\infty \in \mathbf{C}(\mathbb{R}_+, \mathbb{R})$. We now show that

$$\tilde{Y}^\infty = \tilde{Y}. \quad (5.17)$$

By the generalized Skorohod representation theorem (cf. Lemma C.1 of Zhang (2012)), there exists a probability space $(\check{\Omega}, \check{\mathcal{F}}, \check{\mathbb{P}})$, on which stochastic processes $\{\check{Y}^{n'_k}, k \in \mathbb{Z}_+\}$ and \check{Y}^∞ are defined such that

$$\check{Y}^{n'_k} \stackrel{d}{=} \tilde{Y}^{n'_k} \quad \text{and} \quad \check{Y}^\infty \stackrel{d}{=} \tilde{Y}^\infty. \quad (5.18)$$

Moreover, with probability one

$$\check{Y}^{n'_k} \rightarrow \check{Y}^\infty \quad \text{as } n'_k \rightarrow \infty,$$

in the space $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ in the Skorohod J_1 -topology. By Example 12.1 of Billingsley (1999), we also have that $\check{Y}^\infty \in \mathbf{C}(\mathbb{R}_+, \mathbb{R})$. Since the limit is continuous, the convergence (5.16) also holds in the uniform topology. By the Lipschitz continuity of $\Phi_{M,g}$, we have that

$$\Phi_{M,g}(\check{Y}^{n'_k}) \rightarrow \Phi_{M,g}(\check{Y}^\infty) \quad \text{as } n'_k \rightarrow \infty, \quad (5.19)$$

in the space $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ in the uniform topology, hence also in the Skorohod J_1 -topology. In view of (5.18) and Lemma 5.1, we have

$$\Phi_{M,g}(\check{Y}^{n'_k}) \stackrel{d}{=} \Phi_{M,g}(\tilde{Y}^{n'_k}) = \tilde{X}^{n'_k}.$$

Hence, by (5.19),

$$\tilde{X}^{n'_k} \Rightarrow \Phi_{M,g}(\tilde{Y}^\infty) \quad \text{as } n'_k \rightarrow \infty \quad (5.20)$$

in the Skorohod J_1 -topology. So, in view of $\Phi_{M,g}(\tilde{Y}^\infty) \in \mathbf{C}(\mathbb{R}_+, \mathbb{R})$, we know that $\{\tilde{X}^{n'_k}, n'_k \in \mathbb{Z}_+\}$ is C -tight. A direct consequence is the C -tightness of the queue length processes $\{\tilde{Q}^{n'_k}, n'_k \in \mathbb{Z}_+\}$. Hence, (5.17) follows directly from Theorem 2.1 and Proposition 5.1. Therefore, in view of (5.17) and (5.20) we have

$$\tilde{X}^{n'_k} \Rightarrow \Phi_{M,g}(\tilde{Y}) \text{ as } n'_k \rightarrow \infty$$

in the Skorohod J_1 -topology. This completes the proof. \square

6 Conclusion

This paper provides a unified approach to studying the basic queueing model with a single customer class and a single server pool in a wide range of heavy traffic regimes and with a generalized scaling of the patience time. The results in this paper encompass all the previous works summarized in Table 1.1 and extend our understanding of the model to cases that have not yet been explored. Our results generalize those of Atar (2012) to $G/M/\sqrt{n} + GI$, and also allow the hazard rate scaling on the patience time distribution. We also generalize the diffusion approximation of Reed and Tezcan (2011) to $G/GI/n + GI$. Due to technical challenges, the study of the service completion process in the regime parameterized by $\alpha \in (0, 1)$ still remains open.

Acknowledgement

The authors are indebted to Avi Mandelbaum at Technion for helpful discussion, in particular for introducing them to the NDS heavy traffic regime and calling their attention to Rami Atar's paper (Atar (2012)). The authors also thank Rami Atar at Technion for helpful discussion. The research is supported by a start-up grant from NUS Business School and a GRF grant (Project No. 622110) from Hong Kong Research Grants Council.

References

- Atar, R. (2012). A diffusion regime with non-degenerate slowdown. *Oper. Res.*, Forthcoming.
- Baccelli, F. and G. Hebuterne (1981). On queues with impatient customers. In K. F.J. (Ed.), *Performance 81*, pp. 159–179. North-Holland Publishing Company.
- Billingsley, P. (1999). *Convergence of probability measures* (Second ed.). Wiley Series in Probability and Statistics: Probability and Statistics. New York: John Wiley & Sons Inc.
- Boxma, O. J. and P. R. de Waal (1994). Multiserver queues with impatient customers. *ITC 14*, 743–756.
- Chow, Y. S. and H. Teicher (2003). *Probability theory: Independence, interchangeability, martingales*. New York: Springer-Verlag.
- Dai, J. G. and W. Dai (1999). A heavy traffic limit theorem for a class of open queueing networks with finite buffers. *Queueing Syst.* 32(1-3), 5–40.
- Dai, J. G. and S. He (2010). Customer abandonment in many-server queues. *Math. Oper. Res.* 35(2), 347–362.
- Dai, J. G., S. He, and T. Tezcan (2010). Many-server diffusion limits for $G/Ph/n + GI$ queues. *Ann. Appl. Probab.* 20(5), 1854–1890.

- Dudley, R. M. (2002). *Real analysis and probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.
- Garnik, D. and P. Momčilović (2008). Steady-state analysis of a multiserver queue in the Halfin-Whitt regime. *Adv. in Appl. Probab.* 40(2), 548–577.
- Garnett, O., A. Mandelbaum, and M. Reiman (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management* 4(3), 208–227.
- Glynn, P. W. (1990). Diffusion approximations. In *Stochastic models*, Volume 2 of *Handbooks Oper. Res. Management Sci.*, pp. 145–198. Amsterdam: North-Holland.
- Gurvich, I. (2004). Design and control of the $M/M/N$ queue with multi-class customers and many servers. Master’s thesis, Technion.
- Halfin, S. and W. Whitt (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3), 567–588.
- Iglehart, D. L. and W. Whitt (1970a). Multiple channel queues in heavy traffic. I. *Advances in Appl. Probability* 2, 150–177.
- Iglehart, D. L. and W. Whitt (1970b). Multiple channel queues in heavy traffic. II. Sequences, networks, and batches. *Advances in Appl. Probability* 2, 355–369.
- Jelenković, P., A. Mandelbaum, and P. Momčilović (2004). Heavy traffic limits for queues with many deterministic servers. *Queueing Syst.* 47(1-2), 53–69.
- Kaspi, H. and K. Ramanan (2011a). Law of large numbers limits for many-server queues. *Ann. Appl. Probab.* 21(1), 33–114.
- Kaspi, H. and K. Ramanan (2011b). SPDE limits of many-server queues. Technical report, Technion and Carnegie Mellon University.
- Kingman, J. F. C. (1961). The single server queue in heavy traffic. *Proc. Cambridge Philos. Soc.* 57, 902–904.
- Kingman, J. F. C. (1962). On queues in heavy traffic. *J. Roy. Statist. Soc. Ser. B* 24, 383–392.
- Krichagina, E. V. and A. A. Puhalskii (1997). A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Syst.* 25(1-4), 235–280.
- Lee, C. and A. Weerasinghe (2011). Convergence of a queueing system in heavy traffic with general patience-time distributions. *Stochastic Process. Appl.* 121(11), 2507–2552.
- Mandelbaum, A. (2003, September). Notes from a lecture delivered at the workshop on heavy traffic analysis and process limit of stochastic networks. EURANDOM.
- Mandelbaum, A. and P. Momčilović (2012). Queues with many servers and impatient customers. *Math. Oper. Res.* 37(1), 41–65.
- Palm, C. (1937). Etude des delais d’attente. *EricsonTechnics* 5, 37–56.
- Puhalskii, A. A. and J. E. Reed (2010). On many-server queues in heavy traffic. *Ann. Appl. Probab.* 20(1), 129–195.
- Puhalskii, A. A. and M. I. Reiman (2000). The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. in Appl. Probab.* 32(2), 564–595.

- Reed, J. E. (2007). The $G/GI/N$ queue in the Halfin-Whitt regime II: Idle time system equation. Technical report, New York University.
- Reed, J. E. (2009). The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.* 19(6), 2211–2269.
- Reed, J. E. and Y. Shaki (2012). A fair policy for the $g/gi/n$ queue with multiple server pools. Technical report, New York University and Technion.
- Reed, J. E. and T. Tezcan (2011). Hazard rate scaling for the $GI/M/n + GI$ queue. *Oper. Res.*, Forthcoming.
- Reed, J. E. and A. R. Ward (2004, September). A diffusion approximation for a generalized Jackson network with reneging. In *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing*.
- Reed, J. E. and A. R. Ward (2008). Approximating the $GI/GI/1 + GI$ queue with a nonlinear drift diffusion: hazard rate scaling in heavy traffic. *Math. Oper. Res.* 33(3), 606–644.
- Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res.* 9(3), 441–458.
- Ward, A. R. and P. W. Glynn (2003). A diffusion approximation for a Markovian queue with reneging. *Queueing Syst.* 43(1-2), 103–128.
- Ward, A. R. and P. W. Glynn (2005). A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Syst.* 50(4), 371–400.
- Whitt, W. (1980). Some useful functions for functional limit theorems. *Math. Oper. Res.* 5(1), 67–85.
- Whitt, W. (2002). *Stochastic-process limits*. Springer Series in Operations Research. New York: Springer-Verlag.
- Whitt, W. (2003). How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* 51(4), 531–542.
- Whitt, W. (2005). Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.* 30(1), 1–27.
- Whitt, W. (2007). Proofs of the martingale FCLT. *Probab. Surv.* 4, 268–302.
- Williams, R. J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Syst.* 30(1-2), 27–88.
- Zeltyn, S. and A. Mandelbaum (2005). Call centers with impatient customers: Many-server asymptotics of the $M/M/n + G$ queue. *Queueing Syst.* 51(3-4), 361–402.
- Zhang, J. (2012). Fluid models of many-server queues with abandonment. *Queueing Syst.*, Forthcoming.

A A Regulator Mapping

Proof of Lemma 5.1. Note that the special case where $g \equiv 0$ was proved by Reed (2007) (cf. Proposition 7 there). In other words, for any $y \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$, there exists a unique $x \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$ such that

$$x(t) = y(t) + \int_0^t (x(t-s))^- dM(s). \quad (\text{A.1})$$

Define the mapping Φ_M by $x = \Phi_M(y)$, then Φ_M is Lipschitz continuous in the topology of uniform convergence over bounded intervals, measurable with respect to the Borel σ -field generated by the Skorohod J_1 topology. Let Λ_T^M denote the Lipschitz constant of Φ_M in the topology of uniform convergence over the bounded interval $[0, T]$.

In order to deal with the integral equation (5.10), we consider the mapping Ψ_g given by $z = \Psi_g(y)$ with

$$z(t) = y(t) + \int_0^t g((\Phi_M(z)(s))^+) ds \quad \text{for } y \in \mathbf{D}(\mathbb{R}_+, \mathbb{R}). \quad (\text{A.2})$$

Clearly, $x = \Phi_M(\Psi_g(y))$ (that is, $\Phi_{M,g} = \Phi_M \circ \Psi_g$) is a solution to (5.10). Thus, to prove the lemma, it suffices to show that

- (a) the existence and uniqueness of the solution to (A.2);
- (b) Ψ_g is Lipschitz continuous with the topology of uniform convergence over bounded intervals;
- (c) Ψ_g is measurable with respect to the Borel σ -field generated by the Skorohod J_1 -topology.

We focus our analysis on the bounded interval $[0, T]$ for some $T > 0$. Let Λ_T^g be the Lipschitz constant of g on the interval $[0, T]$. Let $\delta = 2/(3\Lambda_T^M \Lambda_T^g)$.

Proof of (a): We first show the existence of a solution. Define $u_0 = 0$ and u_n iteratively by

$$u_{n+1}(t) = y(t) + \int_0^t g((\Phi_M(u_n)(s))^+) ds$$

for all $n \geq 1$. Then

$$u_{n+1}(t) - u_n(t) = \int_0^t [g((\Phi_M(u_n)(s))^+) - g((\Phi_M(u_{n-1})(s))^+)] ds.$$

Now we will show that

$$\|u_{n+1} - u_n\|_{j\delta} \leq j^j n^j \left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} \quad \text{for } j = 1, 2, \dots, \lfloor \delta^{-1}T \rfloor + 1. \quad (\text{A.3})$$

For $j = 1$,

$$\|u_{n+1} - u_n\|_\delta \leq \Lambda_T^M \Lambda_T^g \|u_n - u_{n-1}\|_\delta \times \delta \leq \frac{2}{3} \|u_n - u_{n-1}\|_\delta.$$

Since $g(0) = 0$, we have $\|u_1 - u_0\|_\delta = \|y\|_\delta \leq \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta}$, as a result,

$$\|u_{n+1} - u_n\|_\delta \leq \left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} \leq n \left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta}.$$

Now assume that we have proved (A.3) for $j \leq k$. Then for $j = k + 1$,

$$\begin{aligned} \|u_{n+1} - u_n\|_{(k+1)\delta} &\leq \sum_{j=1}^k \Lambda_T^M \Lambda_T^g \delta \|u_n - u_{n-1}\|_{j\delta} + \Lambda_T^M \Lambda_T^g \delta \|u_n - u_{n-1}\|_{(k+1)\delta} \\ &= \sum_{j=1}^k \frac{2}{3} \|u_n - u_{n-1}\|_{j\delta} + \frac{2}{3} \|u_n - u_{n-1}\|_{(k+1)\delta} \\ &\leq \sum_{j=1}^k \frac{2}{3} j^j (n-1)^j \left(\frac{2}{3}\right)^{n-1} \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} + \frac{2}{3} \|u_n - u_{n-1}\|_{(k+1)\delta} \\ &\leq k^{k+1} n^k \left(\frac{2}{3}\right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} + \frac{2}{3} \|u_n - u_{n-1}\|_{(k+1)\delta}. \end{aligned}$$

Since $\|u_1 - u_0\|_{(k+1)\delta} \leq \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta}$, we have

$$\begin{aligned} \|u_{n+1} - u_n\|_{(k+1)\delta} &\leq k^{k+1} \left(\sum_{i=0}^n i^k \right) \left(\frac{2}{3} \right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} \\ &\leq (k+1)^{k+1} n^{k+1} \left(\frac{2}{3} \right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta}. \end{aligned}$$

Hence, we have proved (A.3), which implies

$$\begin{aligned} \sum_{n=1}^{\infty} \|u_{n+1} - u_n\|_T &\leq \sum_{n=1}^{\infty} \|u_{n+1} - u_n\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} \\ &\leq \sum_{n=1}^{\infty} (\lfloor \delta^{-1}T \rfloor + 1)^{\lfloor \delta^{-1}T \rfloor + 1} n^{\lfloor \delta^{-1}T \rfloor + 1} \left(\frac{2}{3} \right)^n \|y\|_{(\lfloor \delta^{-1}T \rfloor + 1)\delta} \\ &< \infty. \end{aligned}$$

Thus, $\{u_n\}$ is a Cauchy sequence. As $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ is a Banach Space in the uniform metric, the sequence $\{u_n\}$ converges to the limit u^* , which is a solution to (A.2).

The uniqueness of the solution is an immediate consequence of the Lipschitz continuity of Ψ_g , which we prove next.

Proof of (b): For any $y_1, y_2 \in \mathbf{D}(\mathbb{R}_+, \mathbb{R})$, the definition of δ and (A.2) also imply that

$$\|\Psi_g(y_2) - \Psi_g(y_1)\|_{\delta} \leq \|y_2 - y_1\|_{\delta} + \frac{2}{3} \|\Psi_g(y_2) - \Psi_g(y_1)\|_{\delta},$$

Hence, $\|\Psi_g(y_2) - \Psi_g(y_1)\|_{\delta} \leq 3\|y_2 - y_1\|_{\delta}$. Suppose, for $i = 0, 1, \dots, k$,

$$\|\Psi_g(y_2) - \Psi_g(y_1)\|_{i\delta} \leq (3i)^i \|y_2 - y_1\|_{i\delta}. \quad (\text{A.4})$$

We now show that (A.4) holds for $i = k + 1$. For any $t \in [0, (k+1)\delta]$, by the induction assumption and the definition of δ , we have

$$\|\Psi_g(y_2) - \Psi_g(y_1)\|_t \leq \|y_2 - y_1\|_t + \sum_{i=1}^k \frac{2}{3} (3i)^i \|y_2 - y_1\|_t + \frac{2}{3} \|\Psi_g(y_2) - \Psi_g(y_1)\|_t.$$

This implies that (A.4) holds for $i = k + 1$. Continuing the induction till $k = \lfloor \delta^{-1}T \rfloor$ yields the Lipschitz continuity property of Ψ_g .

Proof of (c): Define

$$\Xi(y, u)(t) = y(t) + \int_0^t g((\Phi_M(u)(s))^+) ds.$$

First we prove the function Ξ is measurable with respect to the Borel σ -field generated by the Skorohod J_1 -topology in $\mathbf{D}^2(\mathbb{R}_+, \mathbb{R})$ and $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$. Define $\Pi(y, u)(t) = y(t) + \int_0^t g((u(s))^+) ds$. It is clear that Π is measurable (in fact, continuous) under J_1 topology. Since $\Xi(y, u) = \Pi(y, \Phi_M(u))$ and Φ_M is J_1 -measurable (Proposition 7 of Reed (2007)), the measurability of Ξ is proved.

We know that $\Psi_g(y) = \lim_{n \rightarrow \infty} \Xi^n(y, 0)$, where Ξ^n is iteratively defined by

$$\Xi^n(y, u) = \Xi(y, \Xi^{n-1}(y, u)), \quad n = 1, 2, \dots,$$

with $\Xi^0(y, u) = u$. According to Theorem 2 on page 14 of Chow and Teicher (2003), one can prove by induction that $\Xi^n(y, 0)$ is a measurable function of y for each n . Since Ψ_g is the limit of $\Xi^n(y, 0)$ in the uniform metric on finite intervals, it is also the limit of $\Xi^n(y, 0)$ in J_1 topology. By Theorem 4.2.2 of Dudley (2002), we know that Ψ_g is a measurable function of y with respect to the Borel σ -field generated by the Skorohod J_1 -topology. \square